Center for Science, Technology and Environmental Policy Studies
Arizona State University

cirad
LA RECHERCHE AGRONOMIQUE POUR LE DÉVELOPPEMENT

Institutional and Organizational Factors for Enabling Data Access, Exchange and Use in Genomics Organizations

Eric Welch
Selim Louafi
Federica Fusi
Daniele Manzella

May 2016

The study is jointly funded by the Global Crop Diversity Trust and the Secretariat of the International Treaty on Plant Genetic Resources for Food and Agriculture and is being conducted by researchers at the Arizona State University and CIRAD.

Authors: Eric Welch, Selim Louafi, Federica Fusi, Daniele Manzella

Table of Contents

Executive Summary

In the genomics research fields, there is a growing need for the development of governance principles that encourage use and sharing of materials and data among a wide range of actors in all sectors - government, industry, university and other non-profit - for the development of new knowledge and innovation. Several initiatives and projects are developing information management platforms that aim to be adaptable to exogenous change and sets standards that enable data exchange, integration and interoperability between phenotypic and genotypic data related to genebank holdings.

This study investigated how genomics initiatives and projects have established mechanisms for community building, data sharing and integration and how they addressed competing objectives of partners through institutional structures and organizational designs. It addressed the following questions: What institutions and organizations have been designed to foster public and private returns? What sharing mechanisms and models are evident? What are the tradeoffs to consider?

The study applied a case-based approach. Case study methodology is suitable when the research question is broadly conceived, when complex contextual factors are likely to play a significant role in explaining outcomes, when the research explored an undertheorized contemporary phenomenon, and when appropriate analysis requires multiple sources of evidence (Yin 2011; Stake 1995). The full methodology is presented in the Appendix. Projects and initiatives were identified using a theoretically developed sampling rationale in which selection is based on conceptually justifiable reasons (Glaser & Strauss 1967). Projects and initiatives were selected with the following criteria in mind:

- Scientists and others exchange and use both data and genetic materials.
- Individuals and organizations that access and use materials also contribute knowledge, data and materials back to the program.
- Multiple goals: research, innovation, community building, service.
- Representation from a wide range of disciplines and sub-disciplines of science.
- Diversity of stakeholders and stakeholder institutions including those from public, non-profit and private sectors.
- Involvement in global exchange and global issues, including for developing countries.
- Inclusion of both plant genetic resources (PGR) and other domains such as human health.

Six cases were selected for analysis:

1. Open Science Grid (http://www.opensciencegrid.org)
2. Structural Genomics Consortium (http://www.thesgc.org)
3. iPlant (http://www.iplantcollaborative.org/)
4. Global Alliance for Genomics and Health (https://genomicsandhealth.org)
5. Integrated Breeding Platform (https://www.integratedbreeding.net) and its predecessor Generation Challenge Program (http://www.generationcp.org)
6. International Rice Informatics Consortium (http://iric.irri.org)

The analysis is presented in six sections including: History and drivers; Resource development, aggregation and provision; Membership and heterogeneity; Governance structure, representation, authority and rules; and Sharing approaches.

Key Findings

History and drivers. Each of the projects or initiatives began with a vision, often emanating from a prominent scientist or a small group of scientists, who recognized that existing funding systems, data systems and organizational structures were not responding to new needs related to data and resource intensive research. The vision is typically endorsed by one or more key funders. Key characteristics of the history and evolution of the projects and initiatives studied include: size and scope; primary goal orientation; demand orientation; and niche orientation.

Size and Scope. Projects and initiatives either started with larger size and scope than they eventually implemented or remained small in size while carefully defining scope. Because these initiatives are highly innovative, smaller size and scope give the initiative an opportunity to demonstrate value and effectiveness, reduce complexity and focus on mission.

Primary Goal Orientation. The study identifies three primary goal orientations. In some cases, one or more of these goals overlap.

- A research orientation aims primarily to integrate and organize scientific and technical efforts. Mechanisms used include delivering technical support to already existing research projects; encouraging the community to develop common practices and research methods across projects; or guiding partners towards overarching, common research goals.
- Community-building aims to provide coordination among actors in the field in order to prevent duplication of initiatives, to foster new collaborations and to promote synergies among relevant actors. It can be limited to improving information exchange on existing projects, or it can be more active by brokering connections, services and expertise, or engaging in capacity development activities. Community building often incorporates the development of social capital and trust.
- Service provision consists of sharing IT tools or technical standards to be implemented in single research project. It can extend to the development of a common technical infrastructure.

Demand Orientation. The study identifies two fundamental program design patterns: supply-push and demand-pull. For a supply-push design, decisions about direction and activities are primarily internal with limited *a priori* integration of user perspectives. A demand-pull design focuses on the stated or expressed needs, interests and responses of users. The two designs are end points of a continuum.

The six case studies demonstrate characteristics of both approaches, but there was more evidence of a demand-pull rather than supply-push orientation. Adopting a demand-pull approach requires longer lead time, particularly with dispersed, heterogeneous communities, but it might facilitate commitment towards the goals of the initiative or adoption of the technology. To minimize complexities related to heterogeneity of actors and communities, initiatives may adopt a supply-push approach to speed the initial phases of the initiative, often at the exclusion of critical parties. Supply-push initiatives usually require higher efforts and resources to attract members and disseminate technologies, practices or norms.

Niche orientation. Given the size, scope, goals and design, most projects and initiatives have a clear niche orientation. Projects and initiatives are continuously faced with options for direction of future growth and development, particularly when there is a need to secure sustained funding

and engage heterogeneous stakeholders with diversified needs. Initial niche identification requires clearly defined boundaries of what the organization can (will) and cannot (will not) do. The study shows that decision makers are cognizant of the niche within which their organization operate. Niche determination requires time to specify and energy to maintain, but it also articulates the unique contribution of the project or initiative.

Resource development, aggregation and provision. The resources which the project develops, manages and makes available, represent defining characteristics and catalysts for all cases in this study. It was not always evident at the beginning of each of the projects or initiatives what resources would best address the needs or were in greatest demand. One way to capture the range and concentration of resources offered by the different projects is through the use of a resource framework. For the resource framework, the study identifies six different types of resources: material/data, technical, organizational, institutional, scientific knowledge, and social capital.

Overall, resources form the basis for existence around which the project or initiative is organized. They also help define the scope, design, niche and complexity, discussed in the study. Any single program or initiative provides multiple resources; generally one or two resources are core while others are contingent. Resources can be technically or socially focused. Resources require time to define, put in place and validate. As projects offer more different types of resources, they generally become more complex and more difficult to manage and maintain. Few organizations charge fees for the resources they provide, although some are considering moving in that direction to sustain financial needs.

Membership and heterogeneity. Each project or initiative defines its targeted membership and the extent to which it integrates or not heterogeneous communities and addresses heterogeneous needs. The study identifies three types of heterogeneity among stakeholders in the cases examined:

- Disciplinary heterogeneity, which refers to the diversity of scientific disciplines among stakeholders.
- Sectoral heterogeneity, which refers to diversity stakeholders from public, non-profit and private sectors;
- Geosocial heterogeneity, which refers to the geographical and social diversity of the stakeholders involved.

Because management of heterogeneous communities is difficult, projects generally focus on one or two types of heterogeneity at the expense of the other(s). Incorporation of different types of heterogeneity could lead to fractured leadership or make niche definition of the project or initiative more difficult.

To address significant levels of heterogeneity of actors and communities, projects and initiatives create sub-communities to break heterogeneity into homogeneous groups that facilitate coordination and project effectiveness. Smaller homogenous groups are also able to address collective action challenges more efficiently. But this in turn may create new coordination problems since sub-groups tend to not to collaborate with dissimilar sub-groups, resulting in underachievement of the cooperative potential offered by the initiative. Integration across heterogeneous communities is more effective over time, once the project has already established its functioning rules and structure.

Engaging heterogeneous communities in small projects is a way to show them the positive outcome of joint action. Such projects should be small in scale, should not require actors to invest significant resources and have goals that are achievable in the short-term.

Governance structure, representation, authority and rules. Five dimensions of governance were identified to be critical for the establishment, development and sustainability of projects and initiatives: (1) governance structure, which concerns the position and arrangement of groups and bodies that guide the direction and operation of the project or initiative; (2) source of authority which may derive from representation of interests or competence determined by profession and experience; (3) decision making structure, which may range from highly centralized to highly decentralized; (4) institutions and norms, which comprise rules, policies and procedures; and (5) governance as process.

Governance includes three types of structural components: a high-level and usually independent group of external advisors, a steering committee or board of directors, and a management team led by a single individual, either an executive or a PI. Not all projects or initiatives include all three levels. Projects and initiatives adopt different leadership models including a CEO form in which a leader, usually the initial visionary entrepreneur, continues to be involved over time. But in most cases the leader is described as a coordinating manager who is the head of a small team of individuals with recognized competencies related to a particular scientific, technological or managerial component of the project or initiative. Strong leaders appear to be more likely when an individual serves as the identity of an initiative or when consensus-based decision-making is slow and unproductive. All initiatives have invested considerably in a reliable and efficient management.

Source of authority provides the legitimacy of individuals or groups to make decisions, set strategy and carry out action. Three sources of authority are identified in the case studies: hierarchy, representation and competence. Hierarchy refers to the authority placed in the position of an individual, group or office within an organization. Representation concerns the extent to which relevant stakeholders are either in agreement that their interests are represented by others or are directly involved in decision making and direction. Competence is another form of authority based on recognized knowledge, experience or professional credentials.

Competence-based authority is present in all case studies, mostly in significant degrees. Representation-based decision-making is the strongest in cases where there is a need to establish legitimacy of the organization among members of the community. Ensuring representation can be an obstacle to swift and efficient decision-making. In bodies where different interests hold representation authority, decisions are prone to debate, compromise and synthesis. On the other hand, because representativeness generates trust and buy-in by the membership when decisions are made, long-term implementation may benefit.

Institutions and norms represent key instruments of governance. Institutions and norms include the rules, policies and procedures that are explicitly proscribed or implicitly understood by the leaders and members. Development of institutions and norms is often a particular focus of the projects and initiatives. The institutions that guide projects and initiatives are either tacitly understood or explicitly written down. Early in the lifecycle of a project or initiative, norms may be tacitly understood and their evolution can be tracked by participants who are involved with operations on a day-to-day basis. However, over time, once goals are clear and when there is a

desire to enlarge the membership or a need to communicate new expectations, explicit guidelines and rules may be more important.

Governance processes and patterns. It is clear from the analysis governance systems include formal institutions that guide decisions, interactions and behaviors. Such formal institutions can help diffuse norms and expectations about how people should interact, cooperate, collaborate and share. But governance is also shaped by the actual functions that the initiatives implement and it is the performance and practical application of those functions that actually 'creates' structures, expectations and norms. Very often institutions are created through practice rather than by formal design.

Data sharing approaches. Projects and initiatives adopt a large variety of instruments to support data sharing among members. Instruments are meant to create incentives for data users and contributors and remove potential barriers that might inhibit data sharing. Table 7 (in body of the report) identifies instruments that have been used by projects and initiatives or that have been cited by our interviewees. We rate their effectiveness based on findings from the study and explain additional conditions that might be needed for their effective implementation.

All projects or initiatives analyzed in this study aim to support data-intensive genomics research by promoting data or technical resource sharing across different communities. We classified sharing approaches according to two dimensions. The first one describes whether members are free or not to establish when and under what conditions they want to share their data. Initiatives and projects with "autonomous rules" allow users to set the rules. Initiatives and projects with "common rules" set the rules that members have to follow. The second dimension describes whether the implementation of that data sharing approach requires high technical and organizational resources (including IT infrastructure) or low technical and organizational resources (again, including IT infrastructure). According to those dimensions, we identify four approaches to data sharing: controlled access, data producers, facilitators/brokers and aggregators. Data sharing approaches are often combined within projects and initiatives and can be used to attract different communities. However, when combining different approaches, initiatives and projects must ensure that they have the necessary resources to implement them.

Questions and key tradeoffs. In analyzing the five key areas – drivers, resources, heterogeneity and membership, governance and sharing approach - that need to be considered for managing collaboration in large scale, global genomics projects and initiatives, we highlight important variations across case studies. It is clear that there is no simple or uniform way to address these challenges. Given the complex, interlinked and context-specific nature of this subject, it is not appropriate to recommend prescriptive designs or approaches. Rather, our approach is to direct to consider the tensions that exist between alternatives and suggest useful modes of action. The richness of a comparative analysis precisely lies into the identification of a spectrum of possible actions and directions.

Moreover, by dividing the collaboration challenge into five key areas, we have artificially simplified the overall complexity of collaboration to deeply focus on specific issues one at a time. Nevertheless, we are cognizant that tradeoffs and balances need to be struck across areas. In this section (section 6), we highlight the complexities and linkages by referencing the five areas for three challenging phases for any organization: the formation process; the design of

implementing activities; and the review of critical factors for success. For each, we present a checklist of possible tensions to be considered and we suggest possible modes of action.

Literature review

There are different theoretical frameworks that have conceptualized and analyzed determinants for collaboration and sharing in science. One part of this literature aims at understanding the institutional structure of sharing (i.e. rules, practices, transactions, processes) while another part focuses on the social aspects of sharing (i.e. relationships among individuals, individual skills, competences and motivation). The literature review, which summarize briefly here, has informed our work; allowing us to identify the set of criteria for selection of the case studies and the elements of focus for our analysis. Table 1 summarizes main theoretical approaches that we have reviewed.

The commons framework, developed by Elinor Ostrom in the late '90s for natural resources, is one of the most used approaches to understanding how rules and institutions matter for collective management of pooled resources –fisheries, forests and data repositories. The framework suggests that community sharing can be explained by understanding the dynamics between the shared resource (what is shared), the community (who is sharing) and rules (how resources are shared). Researchers should pay attention to how rights of use and ownership are allocated among actors. We addressed these points in section 2, 3 and 5 of the analysis, by looking at resources mobilized, membership of projects and initiatives and rules for sharing.

Social capital theory introduces the idea that relations among individuals, the way they are structured, the resources that they embed and the common norms and value that regulate them, are significant antecedents for sharing. Research that has applied social capital theory has analyzed how communities develop over time; how trust and reciprocity is formed among actors in a network and how community characteristics, such heterogeneity or homogeneity among actors, influence community outcomes. Social capital is important because it establishes a non-normative basis for sharing and exchanging resources, and facilitates the collective management of common resources. Network theory expands on the idea that social relations matter, predicting that network structure – position of the actors within the network – influences individual and collective outcomes. Network theory also suggests that networks change overtime and that individual and network characteristics influence outcomes simultaneously. We adopt the perspectives offered by social capital theory and network theory to understand how projects and initiatives are designed and how heterogeneity of membership is managed.

Transactional costs theory is used to explain public-private partnerships and how costs and benefits for actors involved play a fundamental role in facilitating or inhibiting collaboration. For collaboration to happen, the governance structure should be able to adapt, coordinate and safeguard exchanges among actors in a more efficient way than any other forms. We saw how the principles of transactional costs theory are indeed applied within public-private initiatives and projects, and how the reduction on transaction costs is a significant incentive for data sharing (section 5).

Finally, collaborative governance theory and management theories aim to explain how effective governance can be created and support overtime, so that projects and initiatives can achieve their goals. Both those approaches focus on how networks, common pools and social relationships can be structured and designed, what resources are necessary to manage them and ensure their sustainability overtime. Collaborative governance theory suggests that interdependence among actors is important and projects or initiatives structure should promote deliberative, consensual and inclusive decision-making processes. The way stakeholders' interests are structured and

represented is important to ensure the ability of the project or initiative to move forward. A strong leadership may be needed to create first relationships, but may be harmful for developing long-term trust across actors. A weak leadership may support trust creation, but might affect project or initiative effective and ability to change and move forward. Compared to collaborative governance theory, management theories examine micro-level elements, such as individual motivation to work together, requisite skills and competences, and group work dynamics. Management theories also look at how initiatives and projects can promote change and how they coordinate daily activities. We address governance and management issues in section 4 of our analysis and we analyze how they are mobilized for data sharing in section 5.

Table 1. Literature review, main theoretical approaches

	Commons framework	Social capital theory	Network theories	Transactional costs theory	Collaborative governance theory	Management theories
Key concepts	Rule and institutions for collective action and management of pooled resources.	Social relationships that explain emergence and functioning of collaboration and exchange.	Effects of different types of relationship structure on individual behavior and collective outcomes	Cost-efficiency considerations for explaining emergence and structure of collaboration	Processes and structures that engage individuals and organizations across sectoral and hierarchical boundaries	Knowledge and human capital management
Constituent elements	• Attributes of the community • Resource characteristics • Rule-in-use • Collective action problem (Social dilemma; public goods)	• Network structure • Common rules and valued • Trust • Reciprocity • Resources hold by actors involved	• Actor position in the network • Flows of information / resources • Structural characteristics of networks • Characteristics of actors in the network	• Costs and benefits analysis of exchange for actors involved • Comparison across collaborative forms	• Actor-centered • Deliberative processes • Legitimacy • Interdependence • Stakeholders and interests	• Skills • Teams • Work design and staffing • IT infrastructure • Rewards • Organizational culture
References	Frischmann, B. M., Madison, M. J., & Strandburg, K. J. (2014) Hess, C. (2012) Ostrom, E. (1990)	Adler, P. S., & Kwon, S.W. (2002) Burt, R. S. (2000) Coleman, J. S. (1988) Nahapiet, J., & Ghoshal, S. (1998)	Ahuja, G., Soda, G., & Zaheer, A. (2012) Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005)	Williamson, O. E. (1979).	Emerson, K., Nabatchi, T., & Balogh, S. (2012). Foss, N. J. (2007) Huxham, C., Vangen, S., Huxham, C., & Eden, C. (2000)	Cabrera, E. F., & Cabrera, A. (2005) Foss, N. J., Husted, K., & Michailova, S. (2010).

Research design

The study followed a case study methodology, which is suitable in cases where the research question is broadly conceived, complex contextual factors are likely to play a significant role in explaining outcomes, the research aims to explore an under-theorized contemporary phenomenon, and when appropriate analysis requires multiple sources of evidence (Eisenhardt & Graebner, 2007; Yin, 2011). Moreover, case study methodology allows deep understanding of the dynamics that take place within a single setting (Eisenhardt, 1989) and eventually comparison of findings across multiple settings (Yin, 2011). In particular, case study methodology fits well in studies where the research question aims to "illuminate a decision or a set of decisions: why they were taken, how they were implemented, and with what results" (Schramm, 1971, p. 4).

For this research project, we adopted a multi-case approach in order to enable exploration of differences within and between cases, and replication of methods across cases to assess commonalities. We compared and contrasted global collaborative initiatives that have different governance structures, participants, goals and rules in practice. Careful case selection is essential under this research scenario to establish similar or contrasting results across cases (Yin, 2011). We identified projects and initiatives using a theoretically developed sampling rationale (Glaser & Strauss, 2012).

The team organized the research in eight phases. The team built sample frame and rationale for case studies selection (phase 1). Selection criteria included factors that were expected to significantly affect the structure and function of large-scale collaborative projects and data sharing initiatives according to the relevant literature analyzed. The team evaluated nineteen cases in the food and agriculture sector and seven cases in the human health sector (phase 2). Based on the initial evaluation, the research team selected ten of the twenty-six cases to conduct exploratory interviews of Project Managers and Executive Directors (phase 3). The interviews focused on governance structure, management processes, exchange and use mechanisms and critical factors for success of each organization. Informed by the interviews, the team selected six cases for in-depth data collection and analysis (phase 4). The research team designed the case study protocol (phase 5) and conducted interviews with staff and "clients", including private sector actors (phase 6). Data collected from the interviews combined with supplementary documentation either provided by the projects or initiatives or collected on their websites were analyzed and compiled into case narratives (phase 7). The same case study protocol was applied across all cases. The research team discussed the findings highlighting similarities and differences across cases and identifying key findings (phase 8). More detail concerning the methodology, sampling procedures and analysis are provided in Appendix 7.

The cases

For the final selection of the cases, the research team followed a maximum variation sampling strategy according to which we "purposefully pick a wide range of variation on dimensions of interest" (Patton, 1990). Dimensions include: engagement of private actor, heterogeneity of actors, involvement of developing countries, participation of members in decision making, data sharing policies and complexity of IT infrastructure.

The selected cases are briefly presented in the following paragraphs. A complete narrative for each case can be found in Appendix 2. iPlant and the Integrated Breeding Platform (IBP) were

selected because of their focus on information technology and community development through shared infrastructure. IBP is also an important case for understanding developing country involvement and evolution of goals, because of its earlier incarnation as the Generation Challenge Program (GCP). The Structural Genomics Consortium (SGC) is a private sector-driven initiative which allows study of company interests and incentives for participation in collaborative initiatives. Open Science Grid (OSG) was selected because of its innovative approach for sharing of computational resources The Global Alliance for Genetic Health (GA4GH) is a relatively large, heterogeneous initiative that aims to design harmonized approaches and policies to promote voluntary, secure and responsible sharing of health and clinical genomics data. Finally, IRIC is the most recently established initiative aiming to develop a data and software resource for the rice research community. It provides means of understanding the early development of an initiative and provides a reference for comparison with other cases.

Case study summaries **(See detailed case descriptions in Appendix 1-6.)**

Open Science Grid is an initiative that facilitates access to distributed computational capacity across research institutions and communities in multiple countries but primarily in the US. Slack computational capacity resources are pooled by a community of volunteers, while OSG tracks capacity availability and matches it with user requests for processing capacity. The initiative is based on autonomy principle, according to which members are free to decide to contribute available capacity or not to the pool, and to establish rules for access and use their resources if desired. Not all users contribute capacity and not all providers of capacity are users.

The **Structural Genomic Consortium** (SGC) is a non-profit organization, supported by pharmaceutical companies, which aims to facilitate joint research activity on a pre-competitive basis. The consortium is committed to undertaking research activities on topics identified by the members of the consortium as critical for new product development but too expensive for any one firm to undertake. The research is undertaken by university scientists using high quality, verifiable methods. Despite its private orientation, the SGC is based on open access principles such that all products and knowledge from its funded research are released into the public domain, without use restrictions.

iPlant is a downloadable, open source data management platform which provides life sciences scientists with informatics tools for the management, analysis, sharing, visualization and cloud storage of large amount of genetic data. The main goals of iPlant are to support data-intensive life science research through the development and deployment of a highly flexible and customizable platform and a user-friendly interface for data management. iPlant supports access to shared databases, but does not pursue a research agenda of its own.

The **Global Alliance for Genetic Health** (GA4GH) is a forum for discussion, design and dissemination of common standards and principles for data sharing. The Global Alliance aims to develop harmonized approaches to data sharing and propose common solutions to data sharing challenges, from technical barriers to security and privacy issues in the field of human genetics and genomics. The work of GA4GH is led by thematic working groups. Solutions and policies developed by working groups are implemented in small demonstration projects to showcase the value of a common approach to data sharing.

The **Integrated Breeding Platform** (IBP) is a data management platform which allows scientists and breeders to manage, analyze and visualize large amount of breeding data. IBP was launched by the Generation Challenge Program (GCP) in response to the increasing need of scientists and breeders, particularly those in developing countries, for bioinformatics tools, and to continue GCP efforts to promote open access to scientific data. IBP offers capacity building programs and encourage the deployment of new technologies to enable traditional and molecular breeding. IBP platforms is still under development and at the current stage it provides minimal facilities for data sharing.

The **International Rice Informatics Consortium** (IRIC) aims to provide a comprehensive repository for rice genetics data by integrating publicly available rice genetics datasets in an easily accessible format. Moreover, IRIC supports rice genetics data production by actively collaborating with key actors in the sector and by facilitating collaboration and communication across the rice research community. The project offers freely available data and data analyses on rice genetics and tools for data analysis and management. The project is oriented towards the integration of genotyping and phenotyping data.

Analysis

The analysis is organized around five key dimensions of each global, collaborative, genomic initiative: history and drivers, resources, community and membership, governance, sharing approaches and policies.

- The history and drivers section analyzes the initial conditions that were necessary for the projects to be launched and developed over time. It examines how the foundational processes and structures were designed and the factors that were critical for success.
- The resources section identifies the resources – including data and material, technical, organizational and institutional resources, knowledge and social capital – that are mobilized by the different projects, where they are located and how members can access them. The analysis attempts to capture the variation of the resources offered by the different projects and initiatives, and to assess production, exchange and use provisions associated with the resources.
- In the community and membership building section, we apply the concept of heterogeneity to recognize diversity among actors and show how initiatives make trade-offs across types of heterogeneity to address coordination and collective action challenges.
- The governance section examines how projects and initiatives are structured, including organizational design, sources of authority, decision making processes, and institutions and norms that regulate project or initiatives activities.
- The sharing approaches section identifies and examines how initiatives promote data sharing, which incentives and rules are designed and which resources are necessary to support data sharing activities.

In all sections, we examine and contrast the case studies to illustrate how choices have been made and how they can be implemented, and we propose some generalizations across cases.

1. History and drivers: critical factors for project foundation and evolution

Projects and initiatives observed at any one point in time are the results of a combination of decisions over time that progressively determine project or initiative membership and governance. Common patterns can be identified across the six cases. Fundamentally, all six cases share a common initial desire to fill institutional, resource or technological gaps in research to bridge increasingly complex yet often isolated communities in ways that enable and accelerate science and innovation. Each project or initiative developed over time moving from a vision-initiated concept to more concrete activities that aim to fill a specific gap. It requires a few years to transition from a broad idea to an actual functioning project.

- The Open Science Grid (OSG) recognized an increasing need for computer processing power among scientists of different disciplines and addressed it by matching requests for processing needs with a network of underutilized processing capacity.
- The Structural Genomics Consortium (SGC) realized the opportunity to broken common yet costly pharmaceutical pre-competitive research interests with university research capacity.
- iPlant recognized the need for data storage, as well as advanced, user-friendly hardware and software tools for data analysis and management in the biological sciences.
- The Global Alliance for Genomics and Health (GA4GH) recognized the need for harmonized approaches to data sharing in the health genomics field in order to reduce technical, scientific and social barriers to data integration, knowledge exchange and learning.
- The Generation Challenge Program (GCP), initially designed to leverage untapped genetic resources for development research, evolved into the Integrated Breeding Platform (IBP) which aims more specifically at the development and diffusion of an integrated breeding management platform for classical and molecular breeding, particularly for development.
- The IRIC envisions a platform to enable the integration and exchange of rice genomic genotyping and phenotypic data to enhance the value of genetic material held in genebanks and catalyze global research.

Each of these organizations began with a vision, often emanating from a prominent scientist or a small group of scientists, recognizing that existing funding systems and organizational structures were not responding to new needs related to the expansion of genomic data and genomic based research. The vision was often endorsed by one or more important funders: iPlant received long-term funding by US NSF; SGC received funding from the Wellcome Trust, although now it is mainly funded by its members; GCP/IBP are mainly funded by the Gates Foundation, the European Commission, and UK Department for International Development; IRIC has yet to receive adequate funding and operates on support from IRRI and GRiSP, and on voluntary cooperative assistance from different research groups; GA4GH also operates on the limited support from research institutions, such as the Ontario Institute for Cancer Research, the University of Cambridge (UK) and the Broad Institute in Massachusetts, and it relies substantially on the voluntary contribution of time and resources of its members.

Empirically, the establishment of these new groups is best characterized as step-wise, evolutionary and fraught with significant challenges that result in reassessment and change of trajectory. Although each case has its own particulars, several generalizations can be made across cases.

Constrained size and scope. Most cases started out with a much broader vision than they were actually able to implement. In part this is due to the range of needs, diversity and number of potential stakeholders, complementary skills required and the ambition of the initiators. Four of the six cases – OSG, SCG, iPlant, GCP/IPB – are relatively mature, meaning that they have existed for several years, while the other two – IRIC and GA4GH – are relatively new. Nevertheless interviews confirmed that the organizations began with a much larger and comprehensive mandate than they currently have and they have consciously sought to maintain focus on a relatively precise scope and mission. Narrowing the scope and mission has allowed those initiatives to effectively design specific activities and move forward in attaining their goals.

This is particularly evident in the case of GCP/IPB. GCP began with the intention to integrate funded research, data production, data sharing and community building across a wide range of stakeholders including actors in economically advanced and developing countries. The highly structured nature of the GCP was retailored into the IBP, which has a much narrower mission, smaller scope and more nimble structure. By starting with only a few companies, SGC was able to demonstrate value and work out the kinks in their complicated, confidential and highly trust-dependent brokering among pharmaceutical companies and between the universities and the companies that make up the membership. SGC has grown incrementally, but the early years were small in scope and size. In a similar sense, while the mission of GA4GH is quite broad, it selects initiatives and projects that are small and narrowly network facilitating, preferring not to enter into competition with members by proposing large grant funded projects.

Sometimes, size and scope reflect the philosophy of the initiators. Interviews with the OSG were clear that it took many years of start-up before they had a recognized service of value. Yet from the beginning, the philosophy was to enable autonomous sharing and facilitate individual use of computer processing. There was no intention to build a new collaborative community. The philosophy of the initiator was also fundamental to the SGC mission. While SGC has procedures to protect companies' data, its initiator has consistently enforced since strict rules concerning the open public access to SGC research outcomes.

> *In general, we find that these projects and initiatives either started with larger size and scope than they eventually implemented or that they have been careful remain small while becoming established. In part, these initiatives are highly innovative, proposing new approaches in a relatively conservative scientific culture. Smaller size and scope gives the initiative an opportunity to demonstrate value and effectiveness, negotiate across multiple interests, and focus on mission.*

Initiative Design - Primary Goal Orientation. We identified three goals that our cases pursue. Some, e.g. SGC, leverage the stock of technological developments and exogenous factors in order to synthesize knowledge and establish research goals aligned with those developments and factors. We defined them as **research-oriented** goals. Some others, e.g. GA4GH, emphasize the iterative aspects of collaboration promoting interactions among community members in order to create favorable conditions for the simultaneous co-production of knowledge at multiple scales

and in multiple locations. We defined them as **community building** goals. Finally, some initiatives, e.g. iPlant, tend to avoid interference with existing collaboration structures (e.g. networks, organizations) and various implicit or explicit rules that are associated with those structures (e.g. on data sharing) and aim, instead, to neutrally provide services to support data-intensive research. These have **service provision** goals.

In each of the three categories, there are important variations that allow projects and initiatives to some goals over others, as shown in the table below.

- A research-oriented approach might include different levels of research goal aggregation. Initiatives can provide technical support to already existing research projects (low level of goal integration) or can support the community in developing common practices and research methods across projects (medium level of goal integration); it can aggregate partners towards overarching, common research goals (high level of goal integration).
- Community-building can materialize in different activities, which we rank in terms of resource intensiveness: exchanging information on existing projects (low level); brokering services and expertise (medium level), and providing capacity development (high level).
- Service provision can consist in different products and services, which again we rank in terms of resource intensiveness: sharing IT tool (low level); adding the deployment of technical standards (medium level); and developing a common technical infrastructure (high level).

Table 2 shows that there are trade-offs among the three goals and, while community building is complementary to all other goals, projects and initiatives tend to create their niche by emphasizing research or service provision goals (see paragraph on Niche Orientation).

Table 2. Primary goals emerged in analyzed cases

	Primary Goals		
	Research	**Community building**	**Service provision**
IBP	0	+++	++(+[1])
GCP	+++	++	0[2]
iPlant	+	++	+++
SGC	+++	++	0
GA4GH	++	++	++
OSG	0	+	+++
IRIC	+[3]	+	++

Research-oriented: (+) collaboration on research projects with matching goal; (++) development of common practices and research methods for enhancing research collaborations; (+++) integration of research goals across all members.
Community building: (+) exchange of information; (++) AND brokerage; (+++) AND capacity building

[1] In future plans IBP will provide a cloud service for storage and exchange of data. At the moment they just provide a standalone platform.
[2] Only storage space
[3] In development

Service provider: (+) software tools; (++) AND standards (e.g. ontologies); (+++) AND common infrastructure

Initiative Design – Supply-Push and Demand-Pull. Two fundamental program design patterns have emerged from the case studies to incentivize engagement for program development: supply-push and demand-pull. The difference between supply-push and demand-pull hinges substantially on the extent to which user needs drive the goals of the initiative. For a supply-push design, decisions about direction and activities of the initiative are primarily internal with limited *a priori* integration of user perspectives. A demand-pull design focuses more on the stated or expressed needs, interests and responses of users. The two designs are end points of a continuum such that any initiative, including the six case studies, demonstrates characteristics of both approaches.

GCP was clearly patterned as a supply-push type of organization in which funding was used to stimulate and orient research for development, data and material sharing and community development. The broad set of goals was established by program designers at the outset, but the diversity of stakeholders, interests and perspectives made it difficult overtime to maintain consistent commitment and compliance with program objectives. IBP continues to be supply-push in that it has *a priori* designed the software, which it offers for no compensation. IBP is responsive to a demand environment seeking breeding software and training but it is focused on disseminating the software rather than co-designing the software with potential users.

OSG is likely the epitome of a demand-pull organization, existing only because of the specific demand from scientists for computer processing time and capacity. There is little attempt by the OSG team to design complementary services or resources as part of the initiative and OSG highly rely on the improvement and integration of already existing software tools. SGC queries the demand environment existing among private company members to identify joint needs. It then brokers an iterative discussion with universities to gain consensus on a joint approach before contracting with universities to supply the research. iPlant is a repository and data management platform with structured guidance for contributors and users, services that are well regarded, but its approach to identifying new services is conservative. While interviewees mentioned other possible avenues for expansion – integration of companies, setting standards for data, data sharing incentives – they were wary of the potential lack of demand.

> *Overall, the cases studies were much more demand-pull rather than supply-push oriented. Interviewees often noted that substantial time during the early stages of the organization was devoted to identifying the precise nature of the existing demand and trying to match it with a viable resource or service that satisfied the demand and did so efficiently and effectively. Adopting a demand-pull approach requires longer lead time, particularly with dispersed, heterogeneous communities but it might facilitate commitment towards the goals of the initiative or adoption of the technology. In order to get around heterogeneity problems, initiatives may be more likely to adopt a supply-push approach to speed the initial phases of the initiative, often at the exclusion of critical parties. Supply-push initiatives usually require higher efforts and resources to attract members and disseminate the technology.*

Niche orientation. Given the aforementioned observations about size, scope and initiative design, it is not surprising that most case interviewees were quick to point out their niche orientation. GA4GH was perhaps the most sophisticated on this front. GA4GH focuses on the needs of the

genomics for human health around three key topics: improving accessibility to genomics data; development of common policies for data sharing; and diffusing best practices across diverse user communities. GA4GH aims to build the connective fascia – technical and institutional – among otherwise differentially connected communities. Their niche secure, they are beginning to expand beyond the US community to connect communities in other countries including developing countries.

One key type of decision frequently identified by interviewees concerned boundary definition. Initiatives are continuously faced with options for direction of future growth and development, particularly when there is a need to secure sustained funding and engage heterogeneous stakeholders with diversified needs. Hence initial niche identification requires clearly defined boundaries of what the organization can (will) and cannot (will not) do. But niche maintenance also requires decision makers to determine over time which are the core activities, interests, resources and services and which are in the periphery.

Indeed, niche identification does not necessarily reduce complexity. Decisions about core services and activities must also consider complementary capacities that are needed to provide contingent or support services. Hence, even the most straightforward of the cases, OSG which matches computational needs with computational capacity, requires substantial technical capacity and a team of over thirty staff to collect, match and process requests from members. Similarly, although iPlant provides a data analysis and management platform, it also trains staff to assist users with the technical advice on data formatting and analysis.

> *In general, case studies showed that decision makers were cognizant of the niche within which their organization operated. Determination of the niche requires time to specify and energy to maintain, but niche articulation specifies the uniqueness and establishes force of attraction of the project or initiative. Filling a niche requires identification of boundaries – decision about what is and is not a function of the initiative – and necessary contingent skills, resources, services and activities that support the core niche.*

2. Resource development, aggregation and provision

The resources which the project develops, manages and makes available, represent defining characteristics and catalysts for all cases in this study. As with other topics discussed in this report, it was not always evident at the beginning of the project or initiative what resources best addressed the needs and were in greatest demand. For example:

- OSG offers the ability to make computational processing resources available. The first few years of its existence were dedicated to developing the network of resource providers and technical and human infrastructure to receive and assign tasks from users.
- SGC spent several years to develop two primary resources that would not otherwise be available: 1) high quality research to discover the 3D structure of human and parasite proteins useful for drug development; 2) established trust to broker and negotiate research foci among competing firms to be conducted by university researchers.
- iPlant was interested in developing a platform to better enable biological research that was increasingly data intensive. Early decisions to focus on software development, standards development for metadata, data storage and some technical assistance related to omics research. Workshops with key researchers were used to inform the design of the platform, the core iPlant resource. As important, iPlant decided not be a collaborator, researcher or broker (although it does provide some introductions among researchers working on similar issues and collaborates on project as technical partner).
- GA4GH generates resources to facilitate the accessibility and integration of genomics data across research communities. Itdevelops policies and standards, testing them through small pilot projects and diffusing best practices based on results. The process through which these resources are developed and diffused strengthens professional networks, thereby enhancing the social capital of the research community
- IBP provides a discrete set of resources including breeding software and training. Additionally, the delivery mechanism is based on a regional hub approach in which champions who are well regarded and knowledgeable about the technology help develop regional communities of users.
- IRIC is at a relatively early stage of development in which it is building its resources – a genomics integrated database on rice and tools to access it. Contributions in time, expertise, data and other resources from established members of the community are critical to IRIC's ability to accomplish its aims.

In all cases, the resources are either developed through initial grant funding or mobilized through voluntary contribution. The pool of resources is usually designed as a joint effort by the project or initiative and the community, linking goals of the visionary, capacity of the funded organization and demand and resources of potential users. Access to the resources is usually provided free of charge (except for SGC) and does not require a matching, equal commitment by users in terms of data sharing, collaboration or other in kind resources.

Resource Framework. One way to capture the range and concentration of resources offered by the different projects is through the use of a resource framework. In this section we define six different types of resources and assess the extent to which the different cases provide resources to users, receive resources from users and link provision with use.

- Material/data resources include genomic or phenotypic data, or genetic materials such as seeds, plant material or DNA. Although data and materials have different properties, we still combine them as a type of input resource to research.
- Technical resources include software (analysis tools, APIs) and human resources for assistance with, by way of example, access and use of existing data or formatting standards and protocols. Technical resources also include equipment, storage space or computational processing time.
- Organizational resources facilitate interaction, collaboration, deliberation, dissemination or some other similar functions among individuals or groups.
- Institutional resources comprise assistance with the development and understanding of legal or regulatory issues or development and dissemination of standards. Data sharing standards can also be included in this category.
- Knowledge resources include scientific collaboration and the knowledge outcomes of collaboration that are embedded within explicit products (journal articles, patents, research protocols) and tacitly understood by scientists.
- Social capital refers to the availability of relational resources – connection, trust, reciprocity and support – within the network of individuals that make up project or initiative members/users.

Given these different resources, projects and initiatives identify, create and offer mixtures that simultaneously address their goals and needs, as shown in table 3. For example, OSG provides mostly technical resources with very few resources in other categories while GA4GH offers multiple types of resources in combination. For instance, only GA4GH provides institutional resources as it aims to establish data standards. By contrast, all the projects or initiatives offer some type of technical resource. This is not surprising as these technical resources lie at the heart of many coordination problems that these cases are designed to address – i.e. data sharing and accessibility.

Most of the projects offer some type of organizational resource that facilitates engagement and interaction among interested parties, although some such as GA4GH offer more than others (iPlant and IRIC). Only SGC and GA4GH are significantly involved in knowledge development and dissemination. Neither IBP nor OSG offer IT infrastructure to share material or data resources, although both IRIC and iPlant are designed to store and make data accessible. Three of the six projects invest few resources in building a community. OSG is fundamentally designed around researcher autonomy, while iPlant and IRIC make weak efforts to actively build user communities.

Table 3. Resources created or offered by the project.

	Resources					
	Material and Data	**Technical**	**Organizational**	**Institutional**	**Knowledge**	**Social Capital**
OSG	0	+++	0	+	0	0
GA4GH	+	++	+++	++	++	+++
SGC	++	+	++	+	+++	+++
IBP	+	+++	++	+	0	+++
iPlant	+++	+++	++	+	+	+
IRIC	+++	++	+	+	+	+

Broadly speaking, resources can be applied in different contexts for different purposes including basic science, innovation or capacity development. Only IBP specifically addresses capacity development, although IRIC has a strong interest in engaging developing country researchers. Most of these resources are designed to operate within advanced R&D settings.

Very few of the organizations studied make requirements on their user base for contributions, financial or otherwise, as shown in table 4. SGC requires payment and contribution of an IP protected target molecule for participation in the research program. iPlant is considering assessing fees for use of its services, but has not made a decision in that regard. Along this line, IBP has established a fee-based system according to which developing country scientists can access to resources for free while developed country scientists and institutions are required to pay a user fee. IRIC does charge a low fee for industry members, but it has only two industry members. In return for financial contributions, IRIC allows industry first access to pre-breeding materials.

Table 4. Requiring contributions from members

	Conditions for access		
	Resources provided linked to resources received	**Charge Fees**	**Other contribution required**
OSG	No	No	No
GA4GH	No	No	No
SGC	Yes	Yes	IPR Target Molecule
IBP	No	Yes, for developed country institutions and researchers	No
iPlant	Under consideration	Under consideration	No
IRIC	Yes, industry only and under consideration for others	Yes, industry only	No

Overall, resources form the basis for existence around which the project or initiative is organized. They also help define the scope, design, niche and complexity, discussed above. Any single program or initiative provides multiple resources; generally one or two resources are core while other contingent resources provide support. Resources can

be technically or socially focused. Resources require time to define, put in place and validate. As projects offer more different types of resources, they generally become more complex and more difficult to manage and maintain. Few organizations charge fees for the resources they provide, although some are considering moving in that direction to sustain financial needs.

3. Membership and heterogeneity

In the first phases, each project or initiative defines its targeted membership and to what extent it integrates or not heterogeneous communities and addresses heterogeneous needs. Building a large and inclusive community is challenging and heterogeneity among actors involved likely affects implementation and sustainability of the project or initiative.

- OSG focuses only on academic scientists, addressing different capacity needs across different disciplines. Actors from same discipline and with homogeneous needs are organized within Virtual Organizations.
- SGC members are exclusively large, international pharmaceutical companies, which are able to pay SGC membership fee. Although SGC allows cross-sectoral collaboration by connecting pharmaceutical companies and universities, all actors involved have a similar capacity level.
- iPlant includes broad, diversified and geographically dispersed research communities in the life sciences field.
- The GA4GH includes a large variety of actors in the field of human genomics, including private, public and nonprofit organizations. Membership is geographically dispersed, although most of members are currently located in OECD countries.
- GCP was based on the collaboration among scientists in nonprofit and public research institutions working on nine selected topics. GCP members were located all around the world and had different capacity.
- IBP includes scientists and breeders in different geographical regions, especially in developing countries, and has developed a regional structure to more closely respond to community needs and issues.
- IRIC focuses only on the rice community, and is trying to progressively include private actors along with public and nonprofit organizations.

Each of these organizations are complex in the sense that: (1) shared knowledge crosses organizational, disciplinary or national boundaries; and (2) actors involved display differences in research practices and methods, capacities, ontologies, human values and epistemologies. To cater for this complexity, we identify three types of heterogeneity among stakeholders present in the projects and initiatives studied and summarized in table 5:

- Disciplinary heterogeneity refers to the diversity of scientific disciplines among stakeholders.
- Sectoral heterogeneity refers to diversity stakeholders from public, non-profit and private sectors;
- Geosocial heterogeneity refers to the geographical and social diversity of the stakeholders involved.

Different types of heterogeneity lead to different collective action and coordination problems. Academic heterogeneity leads to clash of culture and values between infrastructure science (hardware), discovery science (breeders, geneticists) and bio-informatics people (software). Sector heterogeneity leads to a higher risk of free-riding behaviors among actors because of coordination challenges and the need to integrate different sector-based norms values. Finally,

heterogeneity of capacity that is embedded in geosocial diversity, might raise distributional conflicts related to input allocation and resources redistribution.

Table 5. Types of heterogeneity in the analyzed cases

	Heterogeneity		
	Disciplinary Diversity	**Sector Diversity**	**Geosocial Diversity**
GCP	++	+	+++
IBP	+	+(+)	+++
iPlant	++	+	++(+)[4]
SGC	+	++	++[5]
GA4GH	+	+++	++[6]
IRIC	+	+	
OSG	0	+	+

Academic: (+) data-driven (genomics, bio-informatics) science collaboration; (++) data-driven and infrastructure science (material and immaterial) or data-driven and applied science (breeder, clinician); (+++) data-driven, infrastructure science and innovation-driven/applied science collaboration.
Sector: (+) collaboration within one single sector; (++) public-private partnership; (+++) multi-stakeholder partnerships
Country: (+) single country; (++) OECD country; (+++) global

Trade-offs across types of heterogeneity. There are trade-offs that emerge from the coexistence of different types of heterogeneity. It seems difficult to reconcile disciplinary diversity with geosocial diversity. Assigning broad labels to a partnership's purpose gives greater leeway for varied partners to associate and join but can dramatically reduce efficiency and delivery of concrete outputs. For instance, GCP, which aimed to integrate heterogeneous partners in terms of both disciplines and capacity, created a fractured leadership and the project encountered difficulty in establishing and achieving common goals. Without the ability to move forward, the project lost trust from participants and partners whose expectations towards the outcomes of the project were not fulfilled. IBP and GA4GH deal respectively with heterogeneity of capacity and sector, but do so at the expense of other types of heterogeneity. IBP reduced its academic heterogeneity, as it narrowed activities to include only those specifically required for genomic selection and traditional and molecular breeding. GA4GH incorporates sectoral heterogeneity but it does so mainly by incorporating actors with similar capacities.

> *Because management of heterogeneous communities is difficult, projects generally focus on one or two types of heterogeneity at the expense of others. Incorporation of many types of heterogeneity might lead to fractured leadership and the project or initiative might have difficulties in defining its niche, as suggested in the previous section on history and drivers.*

[4] Lead stakeholders and most of collaborators and partners, and users are located in OECD countries. Nevertheless, there are going partnership project in Latin American and there are significant amount of users in China, India and North Africa
[5] SGC has a laboratory in Brazil. Nevertheless, lead stakeholders and interests of the initiative are oriented towards OECD countries.
[6] The initiative is enlarging membership towards African organizations but their number is still extremely marginal.

Managing heterogeneity by creating sub-communities. Even when projects and initiatives select one type of heterogeneity, managing highly heterogeneous communities is difficult. Case studies show that only two initiatives - IBP and GA4GH – are managing high levels of heterogeneity. IBP mainly deals with the capacity heterogeneity challenge that emerges from developing country involvement. GA4GH mainly deals with sector diversity. In some cases, high heterogeneity is managed by creating sub-communities within the project. IBP manages its socio-geographical diversity by creating region-based communities. GA4GH manages heterogeneity through interest-based groups. iPlant, which face a high disciplinary diversity, engage separately with each scientific community. Even OSG, which is the less heterogeneous project, has to address differences in computational capacity requests across disciplines and project size.

> *Projects rarely deal with high heterogeneity. When they do, they tend to create sub-communities to break heterogeneity into homogeneous groups that facilitate coordination and project effectiveness. However, the benefits of this sub-community approach to high heterogeneity are not univocal. Smaller groups could take advantage of their size and homogeneity to address collective action more efficiently, but this in turn may create new coordination problems since subgroups tend to restrict collaboration to their comfort zone and create less incentive for engaging with other sub-groups, hence achieving only a fraction of the possible cooperative potential offered by the initiative.*

Staged approach. The initial aggregation of a single group of actors with common vision and goals is helpful for gathering enthusiasm and energy around the initiative and facilitating the establishment of common rules. SGC started by working with only one private partner. iPlant initially involved only genomics scientists in biology. IBP is building region-based communities by engaging with local institutions on a one-to-one basis. IRIC is focused only on rice genomics researchers and OSG started with a small community of users that had similar computational capacity needs.

Projects can expand their membership basis and increase heterogeneity over time. After consolidation, the initial group can be expanded to include more heterogeneous actors and more diversified needs, as the initiative has already developed its own governance system – structure, decision-making, institutions and norms (see following section 3 of the report). iPlant is engaging with new science communities to create new tools. SGC has progressively enlarged its membership to include the world's largest pharmaceutical companies. OSG expanded its ability to accommodate different computational needs, from small-size to large-size, long-term projects. IRIC and IBP are at an early stage of development, and they are still working toward the integration of heterogeneous communities. IRIC is engaging with private companies but is limited by the fact that its resources are not yet clearly defined. IBP is working towards creating connections across geographically distant communities.

> *Projects choose to start with a single, homogenous community in order to reduce complexity in the design phase of the project. Integration across communities is more effective if it is done over time, once the project has already established its functioning rules and structure.*

Engaging with heterogeneity. Engaging communities on small projects is a way to start building collaboration. Those projects are generally small-sized initiatives that might have a significant

impact on the engagement of heterogeneous groups because they demonstrate the value that can be created by pursuing the organization common goals or applying proposed principles and norms. Small projects could aim to foster coordination among actors in the field to prevent duplication of initiatives, activate economies of scale and promote synergies among relevant actors. GA4GH implements three demonstration projects that need limited funding and pursue short-term goals. The scope of those projects is to showcase the potential of data sharing and integration across datasets. GA4GH uses its demonstration projects as a way to motivate members from different sectors and fields to enhance accessibility to their data and collaborate in the design of harmonized approaches.

Those more specific tasks can be undertaken without necessarily agreeing on the overall goals. The idea is that joint action, even on small activities, should be the beginning of a virtuous circle in which successful action breeds mutual understanding and paves the way for the emergence of joint aims at a later stage (Huxham and Beech, 2003).

> *Engaging heterogeneous communities in small projects is a way to show them the positive outcome of joint action. Projects should be small-scale, not require actors to invest significant resources and have goals achievable in the short-term.*

4. Governance: structure, authority, decision-making and rules

Five dimensions of governance were identified to be critical for the establishment, development and sustainability of projects and initiatives: (1) governance structure, which concerns the position and arrangement of groups and bodies that guide the direction and operation of the project or initiative; (2) source of authority which may derive from representation of interests or competence determined by profession and experience; (3) decision making structure, which may range from highly centralized to highly decentralized; (4) institutions and norms, which comprises rules, policies and procedures that are explicitly proscribed or implicitly understood; and (5) governance as process.

Governance structure includes the existence and arrangement of the committees, groups and units that guide and direct the project or initiative. In general, the structure includes three types of organizations: a management team led by a single individual, either an executive or a PI; a steering committee or board of directors; a high level independent group of external advisors. Not all projects or initiatives include all three levels. As most of the projects and initiatives are science and technology-based, members are mostly scientists or technical specialists, although there are important exceptions.

- OSG is led by a Council which is representative of the different academic fields that currently utilize OSG and the different skills that are required to run the project – technical and scientific skills. Council members are selected and approved by other Council members. A team of PIs/Co-PIs oversees daily project operations and implements decisions taken by the Council.
- SGC is governed by a Board of Directors, which includes a representative for each of the company involved, and supported by two Scientific Committees and an External Scientific Advisory Board. The management of SGC is overseen by a CEO who is responsible for the coordination among companies and between companies and universities, and a project manager who is responsible for the coordination of all SGC internal projects.
- iPlant is led by a PI and a team of Co-PIs. This Executive Team is organized according to function and location, with different functions located at one or two primary sites. Coordination occurs through a series of integrated team meetings – PI/CoPI, function-based, location-based. Most members of iPlant management staff are scientists or technicians. A Science Advisory Board provides scientific oversight and guidance.
- GA4GH is organized around a Steering Committee, composed of representatives who are nominated by GA4GH members. The Executive Director of GA4GH, who is appointed by the Steering Committee, leads a team of five to six Directors who make up the Executive Committee. A separate Strategic Advisory Board is led by an independent chair. The Executive Committee supports and implements activities determined important by the Steering Committee through technical Working Groups which are co-managed by one GA4GH Director and one representative of the Steering Committee.
- IBP is guided by a small Board of Trustees (five to six members), which is in charge of the strategic leadership of the project. Board members include the IBP Director, who is leader of the Management Team. Its predecessor, GCP, was led by a representative board made up of stakeholders who shared authority and responsibility for decision-making.

During the transition from GCP to IBP, GCP established a Scientific and Management Advisory Committee to support the IBP team and provide guidance to the new project.

- The IRRI Director General appoints an IRIC Coordinator who is a full-time staff member of IRRI, with an assistant to manage IRIC activities. IRIC has an Advisory Committee composed of representatives from private sector members, public sector members and IRRI, elected by the IRIC members.

Case studies offered different leadership models including a CEO form in which a leader, usually the initial visionary entrepreneur (SGC and OSG), continues to be involved over time. In one case, GCP, a leader was placed in a strong management role at one point in the lifecycle to streamline operations, make quicker decisions and reduce uncertainty. But in most cases (IRIC, IBP, iPlant, GA4GH) the leader is described as a coordinating manager who is the head of a small team of individuals with recognized competencies related to a particular scientific, technological or managerial component of the project or initiative.

Strong leaders appear to be more likely when an individual serves as the identity of an initiative or when consensus-based decision-making is slow and unproductive. But for the most part, among the cases examined, projects are led by individuals who are respected for their scientific/technical expertise and who are most comfortable directing a team of individuals who are also respected for their own expertise and valued as independent thinkers. It is important to note that most of the cases reviewed are science and technology based and consider their primary constituency to be other members of the science/technology community located primarily, though not exclusively, in OECD countries.

All initiatives have invested considerably in a reliable and efficient management. Interviewees noted that an effective management team is necessary to ensure a high quality product that satisfies demand. Management also guarantees the quality and reliability of internal processes that address sensitive issues such as security and intellectual property. Most initiatives have designated one or two key management personnel who ensure the smooth running of activities.

Source of authority. The governance structure provides a shell that contains the bases of authority necessary to make decisions and motivate action to deliver services, conduct research or build community – the three primary goals discussed in section 1. The source of authority provides the legitimacy of individuals or groups to make decisions, set strategy and carry out action. Authority also provides the basis for monitoring and compliance enforcement to ensure that activities are carried out appropriately.

Three sources of authority are identified in the case studies: hierarchy, representation and competence. Hierarchy refers to the authority placed in the position of an individual, group or office within an organization. Representation concerns the extent to which relevant stakeholders are either in agreement that their interests are represented by others or are directly involved in decision-making and direction. Competence is another form of authority based on recognized knowledge, experience or professional credentials.

Each of these different types of authority are present in the different cases, although hierarchical authority is generally low and has limited importance in organizations that are primarily science and technology based. By contrast representation-based and competence-based authority vary substantially across cases with the general pattern that one of the two forms of authority is more prominent. IBP relies more on competence-based authority to develop the service it provides,

but does not substantially include, for example, breeders from developing countries into the decision making process. Its predecessor, GCP, was quite the opposite, depending heavily on representation-based authority of the different member of the user community. iPlant depends primarily on competence-based authority as does OSG and, to a lesser extent due to the influence of a representative Council, IRIC. The two other organizations, GA4GH and SGC, rely more on representation-based authority, due substantially to the primary goals of the organization – community building – and the importance of social capital as a resource. A summary of sources of authority is presented in table 6.

To emphasize this last point, the source of authority appears to depend substantially on the primary goals of the organization. A service provider may need less representation-based authority than competence based authority, particularly when the product is technical in nature (IBP, OSG and iPlant). A community builder will likely depend on representation-based authority (GA4GH, GCP, SGC).

Table 6. Source of Authority

	Sources of authority	
	Representation	**Competence**
IBP	+	+++
GCP	+++	+
iPlant	+	+++
SGC	+++	++
GA4GH	+++	++
OSG	+	+++
IRIC	++	+++

Representation based authority (level of formal inclusion of relevant stakeholders/users in decision-making): + low; ++ medium; +++ high
Competence based authority (e.g. technical, scientific, strategic): + low; ++ medium; +++ high

> *Competence-based authority is present in all case studies, mostly in significant degrees. Representation-based decision-making is the strongest in cases where there is a need to establish legitimacy of the organization among members of the community. Ensuring representation can be an obstacle to swift and efficient decision-making. In bodies where different interests hold representation authority, decisions are prone to debate, compromise and synthesis. On the other hand, because representation generates trust and buy-in by the membership when decisions are made, long-term implementation may benefit.*

Centralization of decision making. Highly centralized structures enable one person or group to make decisions even if those decisions are relevant for lower levels of the hierarchy or highly distributed units. Decentralized structures allow people or groups at lower levels of the hierarchy or at distributed locations to make decisions.

Centralized decision-making occurs when decision contexts are politicized with different interests, when norms are unsettled or evolving and when actors have variable skill levels. Decentralized decision-making structures are evident when norms are deeply engrained across sites and members such that the norms guide standard approaches to decision rationales; when

decision-makers are highly skilled professionals who make decisions based on professional norms and standards; or when local variation is high creating high communication and information transaction costs. Centralized decision making also occurs when a single product or service is being developed and there is a need to control variability and coordinate the different parts into a whole. Decentralization is more likely to occur when seeking innovative approaches to problem solving that are not likely to be understood well by a central player. Finally, centralization is likely to occur in contexts that include a wide range of different stakeholder interests – where decentralized decision-making will result in the establishment of different priorities and standards depending upon representation. Decentralization is likely to occur in contexts that are dependent on competence-based authority, where collaboration on research is the primary goal.

It is important to note that there is no one best way to organize the decision making structure of a project or initiative. GA4GH, despite its aim to build community through demonstration projects, has a relatively centralized decision-making structure. Centralization in this context allows control over the development of standards and approaches that it produces. IBP follows the same relatively centralized approach. iPlant is more decentralized because of the differences across functional groups and locations. iPlant also benefits from a longer history and the establishment of norms that guide decision making. SGC has a centralized structure concerning strategic decision-making, but it relies on a decentralized structure for decisions concerning research projects, which are carried by different scientific groups at different sites.

Institutions and Norms. It is well understood in the literature and from the cases examined in this study that institutions and norms represent key instruments of governance. Institutions and norms include the rules, policies and procedures that are explicitly proscribed or implicitly understood by the leaders and members. Development of institutions and norms is often a particular focus of the projects and initiatives. For example, GA4GH aims to establish standards for interoperability and principles for data sharing while GCP sought to establish common practices for data sharing. SGC has established norms for coordination, secrecy and public access to knowledge produced by the project.

The institutions that guide projects and initiatives are either tacitly understood or explicitly written down. Early in its lifecycle, a project or initiative institutions may be tacitly understood and their evolution can be tracked by participants who are involved with operations on a day-to-day basis. However, over time, once goals are clear and when there is a desire to enlarge the membership or a need to communicate expectations, explicit guidelines and rules may be more important. . GA4GH, which involve the highest heterogeneity of actors, has developed over time a set of official documents that provide specific indications on data sharing principles that are accepted and shared within its community[7]. IRIC, which is still in its infancy, has yet to engage in an active conversation concerning common data sharing rules and it negotiates rules on a one-to-one basis.

This suggests a possible trade-off between accountability to and accountability for (Agranoff, 2001). An articulated formal institutional design, with formal accountability to (e.g. to a defined

[7] Additional details can be found in the Sharing Approaches section and the GA4GH website concerning the Framework for Responsible Sharing of Genomic and Health-Related Data (https://genomicsandhealth.org/framework) and the GA4GH White Paper (https://genomicsandhealth.org/about-the-global-alliance/key-documents/white-paper-creating-global-alliance-enable-responsible-shar)

membership), may produce unnecessary rigidity in situations where changes to the institutional design (e.g. adding independent scientific oversight) are needed in order to implement new functions (accountability for). While in an initial phase, projects and initiatives often operate based on informal, tacitly understood norms. But once the number and diversity of participating organizations increases, inclusiveness and commitment across heterogeneous actors will likely benefit from transparent governance practices with explicitly stated norms and practices. Two of the six organization (SGC, GA4GH) have developed sophisticated and comprehensive efforts to explicitly articulate the institutions of governance in an effort to build trust and confidence, demonstrate stability, advertise their niche and resources, and ultimately increase membership.

Governance as process. It is clear from the analysis governance systems include formal institutions that guide decisions, interactions and behaviors. Such formal institutions can help diffuse norms and expectations about how people should interact, cooperate, collaborate and share. But governance is also shaped by the actual functions that the initiatives implement and it is the performance and practical application of those functions that actually 'creates' structures, expectations and norms. Very often institutions are created through practice rather than by formal design.

5. Data sharing approaches

All projects or initiatives analyzed in this study aim to support data-intensive genomics research by promoting data or technical resource sharing across different communities. Nevertheless, there is significant variation in how initiatives conceptualize data sharing and which resources they mobilize for the promotion of data sharing among members.

1. OSG supports researcher autonomy as a basis for establishing rules for sharing computational resources. The OSG platform matches capacity needs and capacity availability among users and contributors that have compatible sharing policies.
2. SGC actively engages in data production. Common rules establish that data produced as a result of SGC research activities are freely and publicly accessible online after a 24-month embargo period. However, the SGC ensures that data owned by private companies are privately held and utilized only for internal research purpose.
3. iPlant offers high quality IT infrastructure and storage space which may encourage scientists to share their data, especially those subjected to NSF norms for data publication. It also provides scientists support for data curation and ontology design.
4. GA4GH focuses on data accessibility and brokerage activities by developing common norms and software tools – freely available to the community – that enhance identification or integration of datasets.
5. GCP used funding to leverage and promote sharing.
6. IBP leverages community ties and capacity development efforts to encourage breeders using their software in developing countries to share data. Although the platform does not yet allow easily sharing of data, IBP is developing the potential (and possibly the social capital necessary) for sharing by creating a user community and enabling users to fully utilize the data platform.
7. IRIC is trying to aggregate and integrate publicly available datasets and is collaborating with key actors in the field to produce new data and analyses that will be freely accessible to IRIC community. It also aims to create a high-quality, aggregated rice genomics dataset by combining genotypic and phenotypic data.

All projects and initiatives have designed their data sharing approaches by combining: (1) organizational and technical resources that facilitate data sharing and data sharing related activities – i.e. data curation, formatting and management; and (2) a data sharing framework which includes the rules that allocate rights to use, access and share data and incentives for users and contributors. Sharing approaches are designed to meet expectations of the community, leverage available resources, protect researcher autonomy and encourage data use and contribution.

Instruments for data sharing. Projects and initiatives adopt a large variety of instruments to support data sharing. Instruments are meant to create incentives for data users and contributors and remove potential barriers that might inhibit data sharing. Table 7 shows instruments that have been used by projects and initiatives or that have been cited by our interviewees.

Few of the analyzed projects and initiatives ask members to share their data as a requirement for membership. For instance, the GA4GH membership agreement explicitly states that “members

are not required to commit to a particular threshold of data sharing as a condition of membership in the Global Alliance".[8] IRIC also states that "members are encouraged [not required] to contribute their own data or tools (with clearly defined access rights), thus enriching the globally available data and tools".[9] More specific data sharing requirements might be set forth if members decide to participate in some internal projects. GA4GH specifies that "participation in specific data sharing projects, initiatives, or networks developed or catalyzed by the Global Alliance may carry additional data sharing requirements that go beyond this [membership] agreement". SGC requires members to donate an IPR Target Molecule in order to participate into internal research projects. In general, members are more willing to share data within research projects in order to access to complementary skills, technical expertise and social ties.

Although they do not establish rules for data sharing, projects and initiatives often set guidelines on why and how to share data. GA4GH suggests a Framework for Responsible Sharing of Genomic and Health-Related Data. The Framework provides general guidelines for data sharing in order to establish common values and norms among members. Common values and norms facilitate data sharing by promoting trust-building among parties involved. iPlant, which requires contributors to specify license and terms of use when depositing data in the common repository, suggests to use Creative Commons Universal Public Domain Dedication (CC0).[10] IRIC suggests using the Toronto Agreement.

Table 7. Instrument for data sharing

Instruments for data sharing	Description
Storage space and IT infrastructure	Storage space and IT infrastructure might promote data sharing because they provide users and contributors with access to IT services or data analysis tools.
Access to premium resources / Reduced fee membership	Access to additional resources or reduced membership fee might act as incentives to share data when the community is established and members are aware of the value of membership. Several interviewees refer to the opportunity to relate data contribution to a reduced-price resource access. Nevertheless, none of them has yet adopted this approach.
Embargo to first publication or innovation appropriation (early access to data)	An embargo period on data use allows contributors to capture a return on the investment they have made to produce or collect the data – i.e. for publication or innovation purposes. Projects and initiatives generally recognize contributors with an embargo period on both data they have autonomously produced or data jointly produced in common projects.
Pre-publication / controlled access	Pre-publication or controlled access to data allows researchers to access data before publication or to access sensitive data in a controlled environment. Pre-publication (Toronto Agreement) and controlled access speed scientific discovery without affecting first publication rights of data owners. Pre-

[8] Source: https://genomicsandhealth.org/files/public/Member_Agreement_Organization.pdf
[9] Source: http://iric.irri.org/resources/downloadable-files, Letter of Agreement for Membership – public sector.docs and Letter of Agreement for Membership – private sector.docs.
[10] Source: http://www.iplantcollaborative.org/content/collaboration-policy

	publication and controlled access might be implemented by a third-party organization which monitors whether users and owners rights are respected.
Access to or blockage of funding	Access to funding is a strong motivation for sharing data. NSF policies already require scientists to make their data freely available online. GCP implemented a similar mechanism, providing the last 20% of funding only after projects presented a strategy for sharing data. Nevertheless, this instrument is not effective if not combined with adequate support resources – repository, standard, data quality checks, monitoring or sanctions.
Development of standard for sharing	The development of common standards facilitates sharing by creating trust and shared rules among members of the project or initiative. It also creates conditions for users to access to data and be able to adopt them in further research. Interviewees suggest that this instrument is more effective when it involves relevant communities in standard design.
Facilitating units / organizations	Facilitating units or organizations support data sharing by bridging actors and creating the social structure needed for enabling sharing. GA4GH demonstration projects act as facilitating units integrating interests across different stakeholders and collaborating with them in designing a common system for data sharing. It is important that facilitating teams define and promote equitable sharing among members.
Data pooling	Data pooling might work when all members are required to contribute to the pool and rules are not open for negotiation. Equal contribution to a common pool is applied in the SGC and in demonstration projects in GA4GH. This instrument works best when all actors derive benefits from accessing the pool and actors have equal capacity to contribute to the pool. Data pooling can be combined with other instruments, such as facilitating units.
Access to technical expertise	Given technical skills required for data-intensive research, the provision of technical support to manage and analyze data, or other activities, is likely to be a strong incentive for sharing data. Both iPlant and IRIC are offering to assist members with technical expertise under the condition that data and analyses will be publicly available after the embargo period.
APIs, software tools: interoperability and accessibility	APIs and other software tools promise to solve issues of integration and accessibility of datasets. Most of the analyzed initiatives aims to develop of tools that enable easy integration and access of datasets.
Data quality	High-quality data is an incentive for data use and contribution. Projects and initiatives that offer additional analyses on data or ensure the quality of data provide an added-value service to users. Contributors also benefit from data quality checks as they can take advantage of further data analyses. Initiatives and projects rarelycommit themselves to data quality. Although they try to enhance data access by promoting common standards and formats, they generally discharge responsibility concerning the quality of data hosted in their repository.

Sharing approaches. We classified sharing approaches according to two dimensions. The first captures whether members are free or not to establish when and under what conditions they want to share their data. Initiatives and projects with "autonomous rules" allow users to set the rules. Initiatives and projects with "common rules" set the rules that members have to follow. The second dimension describes the level of resources that are needed to provide incentives to community members to share their data. Based on those dimensions, we separate initiatives and projects into four approaches: (1) facilitators / brokers; (2) controlled access; (3) data aggregators; and (4) data producers.

The four data sharing approaches (Table 8) demonstrate how instruments can be combined in practice to promote data sharing. Initiatives and projects generally apply more than one approach to diversified incentives across communities. When combining different data sharing approaches, initiatives and projects must ensure that they have the necessary resources to implement them.

Table 8. Data sharing approaches

	Low resources	**High resources**
Autonomous rules	Facilitators / Brokers	Aggregators
Common rules	Controlled access	Data producers

Facilitators / Brokers. The initiative or project facilitates data sharing by brokering communication among members and matching potential donors and users with complementary resources. Data sharing occurs among donors and users, often on a dyadic basis; thus, resources are not open nor publicly available to the whole community.

This model promotes data-intensive genomics research by creating a social structure for data sharing and facilitating access to data. The facilitator / broker intermediates among different actors' needs and help them to negotiate common conditions for sharing. As participation is free and there are no rules for sharing data, this model attracts diversified communities. The project or initiative might provide general guidelines or suggest best practices for data sharing.

Facilitators and brokers need few material resources, as they focus on developing relational ties with and within the community to facilitate resource exchange. Social capital – trust, common rules and values – is the key resource (see section 2) because it facilitates data sharing and exchange among members. Facilitators / brokers might offer hardware and software resources to enhance data search. This includes APIs and queries that allow scientists to interrogate multiple datasets in order to locate data or material that they might need. After data are located scientists must negotiate conditions for the exchange with data owners.

Controlled access. An institution, which is considered trustworthy by the members of the community, acts as trustee and allow access to data only for specific research activities and only based on rules controlling of how data is used.

Rules are common across members and are designed by the trustee, according to member requirements. Rules address member concerns – i.e. rights to publication, IP right issues and might include data anonymization, restriction on use or pre-publication conditions. For instance,

researchers might be allowed to examine data and run basic analyses but not to use them for publication or other public purpose. Technical and IT resources provided by initiatives and projects, including queries, APIs or customized systems, provide access to data in a controlled environment.

Benefits for community members include early access to data that advances research in adjacent areas, opportunities to run first analyses while waiting for the embargo to end; access to datasets that provide 'reference sets' for some research fields, because of their broad utility and scale of the project.

SGC is a partial example of controlled access to data that allows collaboration among private actors. The SGC plays trustee's role in the pre-competitive research activities between SGC pharmaceutical companies and universities. SGC guarantees that company data are anonymized and used for research purpose only. Company data cannot be transmitted nor can other companies access them. SGC has created a system to screen anonymized company data libraries to identify data that are useful for current research projects. Only the data found to be useful for the research are accessible to university scientists. Moreover, SGC guarantees that all data discussed in common meetings are presented in aggregated form so that other member companies do not have access to strategically valuable information.

Aggregators**.** The initiative or project aims to aggregate available genomic data by encouraging member contributions (data pooling) or by combining publicly available datasets (aggregation).

Aggregators need to provide strong incentives for members to share their data and for external communities to use them. Projects and initiatives rely both on external incentives – i.e. NSF policies that require funded researchers to make data publicly available – and internal incentives – i.e. access to premium resources, training, or reduced membership fees. Additionally, projects and initiatives in this model provide members with a platform where they can share and access data. The quality, user-friendliness and reputation of the platform and the management team are key to encourage members to upload their data and use data provided. Platforms also facilitate the adoption of common standards and facilitate integration across datasets, through APIs and other software tools.

In this context, rules are autonomously established and designed by data owners. Aggregators usually suggest or adopt a globally recognized data sharing policy, such as the Creative Commons Universal Domain Dedication (CC0) or the Toronto Agreement. Users are free to apply them or implement different policies.

Producers. Community members and the project or initiative collaborate to produce or collect genomics data. Members are willing to collaborate and accept common data sharing rules in exchange to access to the project or initiative resources, such as research funding, technical expertise or complementary skills.

Rules are generally set before the beginning of the collaboration. In well-established projects and initiatives, such as SGC and iPlant, rules are non-negotiable and members or collaborators cannot modify them. Non-negotiable rules are easier to implement, because they provide a common ground for all members and collaborators. Rules protect collaborators' investment in the research activity by establishing an embargo period for them to publish first or to capitalize innovation. At the end of the embargo period - from 12 to 24 months - data are made publicly

and freely available. Additional rules might ensure the quality of data produced and released, through evaluation by independent Advisory Committees (SGC) or release of further analyses that demonstrate data value (IRIC). Quality rules are important because they help the project or initiative builds its reputation in the community.

The producer approach requires substantial resources to support data collection and/or production. Initiatives and projects have different ways to collect financial resources needed: from membership fees to grants and contracts. The lack of adequate financial resources might be a significant constraint for implementing this model. To further encourage sharing, data producers might offer a repository space where common data can be stored and accessed.

Data sharing approaches for all cases appear in table 9.

Table 9. Data sharing approaches in analyzed cases

	Data sharing framework			Data sharing support		
	Shared data	**Rules for sharing**	**Incentives**	**IT infrastructure**	**Technical and organizational resources**	**Data sharing approach**
IBP	GCP nine-crop datasets	None (Autonomous)	Training, access to free resources, free membership, capacity and community building	Data management and analysis platform	Development of customized datasets for different data Technical support for platform use	Aggregator
iPlant	Datasets shared by users	Autonomously decided by contributors (Creative commons)	NSF and other funding agencies policies that require data to be published. High quality, user friendly infrastructure	Data management and analysis platform. Cloud space for common data. APIs, software tools, etc…	Shared development of ontologies with scientific communities. Ontology experts. Technical support for data curation	Aggregator
GA4GH	Data shared by users in demonstration projects	Negotiated among members	Demonstration projects	APIs, software tools	Community-driven working groups and demonstration projects Harmonized policies and approached for responsible sharing	Facilitator / Broker
OSG	Computational capacity	Highly customized at the individual level	Utilization of slack resources	Common infrastructure	Technical support.	Facilitator / Broker

IRIC	3,000 Rice Genome Project data Public data (formatted and curated) Data from collaborative projects	Toronto agreement	Collaboration on project, access to analysis, resources	APIs, software tools, storage space	Technical support. Cooperation on projects Data curation	Data pool Aggregator Producer
SGC	Internally: anonymized data on proteins and compounds Externally: research results	Enforced common rules	Participation to the project; early access to results (24 months embargo)	None	Project management Quality check and data curation	Producer / Controlled access
GCP	Data from funded-projects	Enforced common rules. Mandatory contribution from all members.	Funding for research	Storage space	Common ontologies	Producer

6. Key questions and trade-offs

This study examined five key areas – drivers, resources, heterogeneity and membership, governance and sharing approach – that need to be addressed when establishing large scale genomics projects and initiatives. Analysis highlighted important similarities and differences across cases, but also demonstrated that there are alternative, context specific ways to effectively integrate all five aspects. Of course, by dividing the collaboration challenge into five key areas, we have artificially simplified the overall complexity of collaboration to focus deeply on specific issues one at a time. Nevertheless, this approach has allowed us to show tradeoffs that are created because decisions in one area affect options in the other four areas. In this section, we try to highlight this complexity by looking simultaneously at the five areas during three challenging phases for any organization: formation, implementation and review for continued success. For each, we present a checklist of possible tensions to be considered.

Formation stage: Assessing the initial context

Questions. The main questions to ask here relate to the assessment of the initial conditions that prevail at the start of a project or initiative. Initial conditions are assessed relative to the intended goals, needs, relationships, resources and institutions. Questions include:

a. What extent level of agreement or consensus exists among communities/actors about the directions at the outset of the project or initiative? Have the benefits and contributions been made explicit and are they shared among communities/actors?

b. What is the level of heterogeneity that exists in the community? How different are the members in terms of discipline, sector, capacity, profession, etc.?

c. What social relationships exist among potential participants at the outset? To what extent is there trust among members of the community?

d. What resources does the initiative want to attract and/or provide? Where are they located and how should they be made available?

e. What opportunities and challenges are in place for data sharing? What incentives for data sharing exist and to what extent are they common for all members of the community?

Observations and suggestions

Answers to these questions provide an aid to determining a niche for a program or initiative, but they can also expose trade-offs and challenges. For example, addressing community needs requires the definition of a community boundary which will also determine the participants.

Communities that have broad global missions are likely to include heterogenous partners. If potential participants are already connected, then there is a lower need for establishing relationships. In all cases it is important to set goals that address needs, but goal consensus requires greater time and energy to build when the community is more heterogeneous and less connected. This is because while all participants may share some converging interests, heterogeneity and low connectivity can lead to mistrust and reluctance to fully engage in the new initiative. In this case, it is important to establish a legitimate and accountable governance system that builds trust and willingness to commit time and resources.

Evidence from the case studies also shows that the choice of the resources that will be primarily aggregated in the initiative strongly matters. The accent could be put on the material and technical resources, including data such as the case of iPlant and IBP; on knowledge resources such as in SGC; or on institutional and social capital resources to facilitate interactions, deliberations and build of trust and mutual understanding, as in the case of GA4GH. The three approaches could be

combined over time but experience shows that it is probably unrealistic to deal with all simultaneously.

The choice of what resources to mobilize also influences the choice of the data sharing approach that the initiative or project might pursue. A data producer approach generally requires high organizational, knowledge and technical resources, whereby a facilitator/broker approach requires less investment in material resources but a stronger focus on developing social relationships. As in the initiatives or projects studied, assessing external opportunities, including funding, research niches or agency policies, might generate incentives for data sharing.

Implementation stage of the project or initiative

Questions. The main questions to ask in the implementation phase relate to the governance and management of the community. They include challenges that heterogeneity poses for structures, rules, capacities, research practices and methods, ontologies, values and epistemologies. Questions include:

a. How are the sources the legitimacy, flexibility and accountability mobilized to accomplish the goals? Does the governance structure support the actions and processes of the project or initiative?

b. How can mechanisms be established to effectively integrate heterogeneity?

c. How can social relationships be built and strengthened over time?

d. How should the resource mix be aggregated, produced and made accessible in a sustainable manner over time?

e. How can rules and resources support data sharing processes?

Observations and suggestions

Clarity of governance structures and the actual exercise of decision-making within those structures is important for maintaining commitment and interest of actors. For instance, representation-based authority emanating from a diverse community requires to structure governance in ways that build confidence and participation.

Evidence from the case studies also shows that adjustments of formal structure have taken place following the refinement of goals and evolution of functions. This evolutionary dimension, when applied to governance, suggests that excessively formal structures established at the outset may produce unnecessary rigidity in situations where changes to the institutional design (e.g. adding independent scientific oversight) are needed. While in an initial phase, implementation can be achieved through some level of collegiality and informal norms, once expansion to a larger number and greater diversity of participating organizations occurs, inclusiveness and commitment across heterogeneous actors is better maintained though transparent governance practices.

Three distinct strategies to deal with heterogeneity emerge from the case studies: the first is limit heterogeneity initially while including more diverse actors as the initiative evolves over time; the second is to create homogeneous sub-groups; and the third is to engage upfront with heterogeneity but only on small-scale project-basis as a proof of concept.

Interactions among actors could gradually establish a pattern of data sharing. Interactions may be intermediated by the technical infrastructure, formal or informal rules, or some form of collaboration. Resources that support the implementation of data sharing approaches are critical during the implementation phase to demonstrate in the short run the project effectiveness and retain first participants. Starting small, with a homogenous group might be an effective, proof-of-concept strategy.

Trajectory and performance feedback

Questions. The main questions to ask in this phase relate to the how well the project or initiative is taking hold and proceeding in the intended direction, and whether early corrective action is needed. The aim is not specifically evaluative, rather it offers an early mechanism for guidance and smoothing the integration of the five areas and to ensure that the intended trajectory is not fundamentally altered. Questions include:

a. Are goals integrated into members' activities? Are goals recognized within the community? Is commitment to goals widespread?
b. Have the mechanisms for managing heterogeneity overcome collective action/coordination problems and produced concrete outputs?
c. Has the network expanded and have new collaborations been established? Have perceptions of trust increased among groups?
d. Do resources adequately support core and periphery activities of the organization, across all members?
e. What is the progress on data sharing? What barriers still exist for data sharing? Do selected incentives induce and enable data sharing?

Observations and suggestions

Answering these questions is helpful to monitor the progress and ensure that adequate corrective actions are taken early on. For example, case studies show that early commitment of community members to the goals of the initiative is likely a good indicator of longer term sustainability of the initiative. Similarly, feedback about management and inclusion of heterogeneity and levels of trust may help inform whether collective action or data sharing are likely. The key question about data sharing is whether the chosen approach has been able to accommodate the needs and interests of targeted actors, particularly those from particular sub-communities such as industry or developing countries. At this stage, it is important to think about further instruments that might be designed to promote engagement with other relevant communities, or expand and revise goals and activities as needed.

Table 10. Key questions and trade-offs

Phase 1 Assessing the context	Phase 2 Implementing the initiative	Phase 3 Trajectory & Performance Feedback
Assessing shared goals	**Designing governance for shared goals**	**Goal integration feedback**
What extent level of agreement or consensus exists among communities/actors about the directions at the outset of the project or initiative? Have the benefits and contributions been made explicit and are they shared among communities/actors?	How are the sources the legitimacy, flexibility and accountability mobilized to accomplish the goals? Does the governance structure support the actions and processes of the project or initiative?	Are goals integrated into members' activities? Are goals recognized within the community? Is commitment to goals widespread?
Assessing heterogeneity	**Integrating heterogeneity**	**Heterogeneity feedback**
What is the level of heterogeneity that exists in the community? How different are the members in terms of discipline, sector, capacity, profession, etc.?	How can mechanisms can be established to effectively integrate heterogeneity?	Have the mechanisms for managing heterogeneity overcome collective action/coordination problem and produced concrete outputs?
Assessing social structure and trust	**Leveraging the social structure**	**Relational feedback**
What social relationships exist among potential participants at the outset? To what extent is there trust among members of the community?	How can social relationships be built and strengthened over time?	Has network expanded and have new collaborations been established? Have perceptions of trust increased among groups?
Assessing resource mix	**Creating the resource mix**	**Resource accessibility feedback**
What resources the initiative wants to attract and where they are located? Where are they located and how should they be made available?	How should the resource mix be aggregated, produced and accessible to be sustainable over time?	Do resources adequately support core and periphery activities of the organization, across all members?
Assessing opportunities for data sharing	**Incentivizing and regulating data sharing**	**Data sharing effectiveness**

What opportunities and challenges exist for data sharing? What incentives for data sharing exist and to what extent are they common for all members of the community?	How can rules and resources support data sharing processes?	What is the progress on data sharing? What barriers still exist for data sharing? Do selected incentives induce and enable data sharing?

References

Adler, P. S., & Kwon, S.-W. (2002). Social capital: Prospects for a new concept. *Academy of Management Review*, *27*(1), 17–40.

Agranoff, R., & McGuire, M. (2001). Big Questions in Public Network Management Research. *Journal of Public Administration Research and Theory*, *11*(3), 295–326.

Ahuja, G., Soda, G., & Zaheer, A. (2012). The Genesis and Dynamics of Organizational Networks. *Organization Science*, *23*(2), 434–448. http://doi.org/10.1287/orsc.1110.0695

Baxter, P., & Jack, S. (2008). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The Qualitative Report*, *13*(4), 544–559.

Bouffard, M., Phillips, M. S., Brown, A. M. K., Marsh, S., Tardif, J.-C., & van Rooij, T. (2010). Damming the genomic data flood using a comprehensive analysis and storage data structure. *Database: The Journal of Biological Databases and Curation*, *2010*. http://doi.org/10.1093/database/baq029

Burt, R. S. (2000). The Network Structure Of Social Capital. *Research in Organizational Behavior*, *22*, 345–423. http://doi.org/10.1016/S0191-3085(00)22009-1

Cabrera, E. F., & Cabrera, A. (2005). Fostering knowledge sharing through people management practices. *The International Journal of Human Resource Management*, *16*(5), 720–735. http://doi.org/10.1080/09585190500083020

Chokshi, D. A., Parker, M., & Kwiatkowski, D. P. (2006). Data sharing and intellectual property in a genomic epidemiology network: policies for large-scale research collaboration. *Bulletin of the World Health Organization*, *84*(5), 382–387. http://doi.org/10.1590/S0042-96862006000500018

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, *94*, S95–S120.

Contreras, J. L. (2014). *Constructing the Genome Commons* (SSRN Scholarly Paper No. ID 2474405). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2474405

Dedeurwaerdere, T. (2006). The institutional economics of sharing biological information. *International Social Science Journal*, *58*(188), 351–368. http://doi.org/10.1111/j.1468-2451.2006.00623.x

Dove, E. S., Faraj, S. A., Kolker, E., & Özdemir, V. (2012). Designing a post-genomics knowledge ecosystem to translate pharmacogenomics into public health action. *Genome Medicine*, *4*(11), 91. http://doi.org/10.1186/gm392

Doz, Y. L., Olk, P. M., & Ring, P. S. (2000). Formation processes of R&D consortia: which path to take? Where does it lead? *Strategic Management Journal*, *21*(3), 239–266. http://doi.org/10.1002/(SICI)1097-0266(200003)21:3<239::AID-SMJ97>3.0.CO;2-K

Eckersley, P., Egan, G. F., Schutter, E. D., Yiyuan, T., Novak, M., Sebesta, V., … Toga, A. W. (2003). Neuroscience data and tool sharing. *Neuroinformatics*, *1*(2), 149–165. http://doi.org/10.1007/s12021-003-0002-1

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Academy of Management Review*, *14*(4), 532–550. http://doi.org/10.5465/AMR.1989.4308385

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory Building from Cases: Opportunities and Challenges. *The Academy of Management Journal*, *50*(1), 25–32. http://doi.org/10.2307/20159839

Elta Smith. (n.d.). Ownership and responsibility: public property in Creative Commons and rice genomics.

Emerson, K., Nabatchi, T., & Balogh, S. (2012). An Integrative Framework for Collaborative Governance. *Journal of Public Administration Research and Theory*, *22*(1), 1–29. http://doi.org/10.1093/jopart/mur011

Foss, N. J. (2007). The Emerging Knowledge Governance Approach: Challenges and Characteristics. *Organization*, *14*(1), 29–52.

Foss, N. J., Husted, K., & Michailova, S. (2010). Governing Knowledge Sharing in Organizations: Levels of Analysis, Governance Mechanisms, and Research Directions. *Journal of Management Studies*, *47*(3), 455–482. http://doi.org/10.1111/j.1467-6486.2009.00870.x

Foster, M. W., & Sharp, R. R. (2007). Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nature Reviews Genetics*, *8*(8), 633–639. http://doi.org/10.1038/nrg2124

Frischmann, B. M., Madison, M. J., & Strandburg, K. J. (2014). *Governing Knowledge Commons*. Oxford University Press on Demand. Retrieved from https://books.google.com/books?hl=en&lr=&id=0UOXBAAAQBAJ&oi=fnd&pg=PP1&dq=governing+knowledge+commons&ots=OZc-BN4jsV&sig=6c3V_lFnuhSYYaZveZVujKfxESY

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Publishing Company.

Herbert J., R., & Irene S., R. (2004). *Qualitative Interviewing: The Art of Hearing Data* (3rd ed.). Sage Publications.

Hess, C. (2012). The Unfolding of the Knowledge Commons. *St Antony's International Review*, *8*(1), 13–24.

Huxham, C., Vangen, S., Huxham, C., & Eden, C. (2000). The Challenge of Collaborative Governance. *Public Management Review*, *2*(3), 337–358. http://doi.org/10.1080/14719030000000021

Huxham, C., Vangen, S., Huxham, C., & Eden, C. (2000). The Challenge of Collaborative Governance. *Public Management Review*, *2*(3), 337–358. http://doi.org/10.1080/14719030000000021

Knoppers, B. M. (2009). Genomics and policymaking: from static models to complex systems? *Human Genetics*, *125*(4), 375–379. http://doi.org/10.1007/s00439-009-0644-7

Kosseim, P., Dove, E. S., Baggaley, C., Meslin, E. M., Cate, F. H., Kaye, J., … Knoppers, B. M. (2014). Building a data sharing model for global genomic research. *Genome Biology*, *15*(8), 430. http://doi.org/10.1186/s13059-014-0430-2

Möllering, G. (2005). *Understanding Trust from the Perspective of Sociological Neoinstitutionalism: The Interplay of Institutions and Agency* (SSRN Scholarly Paper No. ID 1456838). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1456838

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review*, *23*(2), 242–266.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Patton, M. (1990). Purposeful sampling. In *Qualitative evaluation and research methods* (pp. 168–186). Beverly Hills (CA): Sage.

Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005). Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences. *American Journal of Sociology*, *110*(4), 1132–1205. http://doi.org/10.1086/421508

Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005). Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences. *American Journal of Sociology*, *110*(4), 1132–1205. http://doi.org/10.1086/421508

Schramm, W. (1971). Notes on Case Studies of Instructional Media Projects. Retrieved from http://eric.ed.gov/?id=ED092145

Schroeder, M. P., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). Visualizing multidimensional cancer genomics data. *Genome Medicine*, *5*(1), 9. http://doi.org/10.1186/gm413

Stake, R. E. (1995). *The art of case study research*. Sage Publications.

Strandburg, K. J., Frischmann, B. M., & Cui, C. (2014). The Rare Diseases Clinical Research Network and the Urea Cycle Disorders Consortium as nested knowledge commons. In B. M. Frischmann, M. J. Madison, & K. J. Strandburg (Eds.), *Governing Knowledge Commons* (pp. 155–207). Oxford University Press.

Tasselli, S., Kilduff, M., & Menges, J. I. (2015). The Microfoundations of Organizational Social Networks A Review and an Agenda for Future Research. *Journal of Management*, 0149206315573996.

Tsai, W., & Ghoshal, S. (1998). Social Capital and Value Creation: The Role of Intrafirm Networks. *The Academy of Management Journal*, *41*(4), 464–476. http://doi.org/10.2307/257085

Welch, E., & Louafi. (n.d.). Contested Inputs for Scientific Research: Why Access to Biological Materials Is Blocked.

Williamson, O. E. (1979). Transaction-Cost Economics: The Governance of Contractual Relations. *The Journal of Law & Economics*, *22*(2), 233–261.

Yin, R. K. (2011). *Applications of case study research*. Sage. Retrieved from https://books.google.com/books?hl=en&lr=&id=FgSV0Y2FleYC&oi=fnd&pg=PR1&dq=yin+case+studies+sage&ots=41b0MpqgVp&sig=QADGO_RPpnDz-93c02oqa11BdPI

References from literature review

Adler, P. S., & Kwon, S.-W. (2002). Social capital: Prospects for a new concept. *Academy of Management Review*, *27*(1), 17–40.

Ahuja, G., Soda, G., & Zaheer, A. (2012). The Genesis and Dynamics of Organizational Networks. *Organization Science*, *23*(2), 434–448. http://doi.org/10.1287/orsc.1110.0695

Benefits and Best Practices of Rapid Pre-Publication Data Release. (2009). *Nature*, *461*(7261), 168–170. http://doi.org/10.1038/461168a

Burt, R. S. (2000). The Network Structure Of Social Capital. *Research in Organizational Behavior*, *22*, 345–423. http://doi.org/10.1016/S0191-3085(00)22009-1

Cabrera, A., & Cabrera, E. F. (2002). Knowledge-Sharing Dilemmas. *Organization Studies*, *23*(5), 687–710. http://doi.org/10.1177/0170840602235001

Chokshi, D. A., Parker, M., & Kwiatkowski, D. P. (2006). Data sharing and intellectual property in a genomic epidemiology network: policies for large-scale research collaboration. *Bulletin of the World Health Organization*, *84*(5), 382–387. http://doi.org/10.1590/S0042-96862006000500018

Colledge, F., Elger, B., & Howard, H. C. (2013). A Review of the Barriers to Sharing in Biobanking. *Biopreservation and Biobanking*, *11*(6), 339–346. http://doi.org/10.1089/bio.2013.0039

Colledge, F., Elger, B., & Howard, H. C. (2013). A Review of the Barriers to Sharing in Biobanking. *Biopreservation and Biobanking*, *11*(6), 339–346. http://doi.org/10.1089/bio.2013.0039

Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. *Science*, *300*(5617), 286–290. http://doi.org/10.1126/science.1084564

Constable, H., Guralnick, R., Wieczorek, J., Spencer, C., Peterson, A. T., & The VertNet Steering Committee. (2010). VertNet: A New Model for Biodiversity Data Sharing. *PLoS Biol*, *8*(2), e1000309. http://doi.org/10.1371/journal.pbio.1000309

Cook-Deegan, R. (2006). The science commons in health research: structure, function, and value. *The Journal of Technology Transfer*, *32*(3), 133–156. http://doi.org/10.1007/s10961-006-9016-9

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science*, *41*(5), 667–690. http://doi.org/10.1177/0306312711413314

Elta Smith. (n.d.). Ownership and responsibility: public property in Creative Commons and rice genomics.

Emerson, K., Nabatchi, T., & Balogh, S. (2012). An Integrative Framework for Collaborative Governance. *Journal of Public Administration Research and Theory*, *22*(1), 1–29. http://doi.org/10.1093/jopart/mur011

Foss, N. J. (2007). The Emerging Knowledge Governance Approach: Challenges and Characteristics. *Organization*, *14*(1), 29–52.

Foster, M. W., & Sharp, R. R. (2007). Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nature Reviews Genetics*, *8*(8), 633–639. http://doi.org/10.1038/nrg2124

Frischmann, B. M., Madison, M. J., & Strandburg, K. J. (2014). *Governing Knowledge Commons*. Oxford University Press on Demand. Retrieved from

https://books.google.com/books?hl=en&lr=&id=0UOXBAAAQBAJ&oi=fnd&pg=PP1&dq=governing+knowledge+commons&ots=OZc-BN4jsV&sig=6c3V_lFnuhSYYaZveZVujKfxESY

Giest, S., & Howlett, M. (2014). Understanding the pre-conditions of commons governance: The role of network management. *Environmental Science & Policy*, *36*, 37–47. http://doi.org/10.1016/j.envsci.2013.07.010

Halewood, M. (2013). What kind of goods are plant genetic resources for food and agriculture? Towards the identification and development of a new global commons. *International Journal of the Commons*, *7*(2), 278–312.

Hess, C. (2012). The Unfolding of the Knowledge Commons. *St Antony's International Review*, *8*(1), 13–24.

Hess, C., & Ostrom, E. (2003). Ideas, Artifacts, and Facilities: Information as a Common-Pool Resource. *Law and Contemporary Problems*, *66*(1/2), 111–145.

Hess, C., & Ostrom, E. (2006). A framework for analysing the microbiological commons. *International Social Science Journal*, *58*(188), 335–349. http://doi.org/10.1111/j.1468-2451.2006.00622.x

Hess, C., & Ostrom, E. (n.d.). Mertonianism Unbound? Imagining Free, Decentralized Access to Most Cultural and Scientific Material. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6284190

Huysman, M., & Wulf, V. (2006). IT to support knowledge sharing in communities, towards a social capital analysis. *Journal of Information Technology*, *21*(1), 40–51.

Including, N. C. D. S. for A. D. T. S. T., Perrino, T., Howe, G., Sperling, A., Beardslee, W., Sandler, I., … Brown, C. H. (2013). Advancing Science Through Collaborative Data Sharing and Synthesis. *Perspectives on Psychological Science*, *8*(4), 433–444. http://doi.org/10.1177/1745691613491579

Including, N. C. D. S. for A. D. T. S. T., Perrino, T., Howe, G., Sperling, A., Beardslee, W., Sandler, I., … Brown, C. H. (2013). Advancing Science Through Collaborative Data Sharing and Synthesis. *Perspectives on Psychological Science*, *8*(4), 433–444. http://doi.org/10.1177/1745691613491579

Joly, Y., Dove, E. S., Knoppers, B. M., Bobrow, M., & Chalmers, D. (2012). Data Sharing in the Post-Genomic World: The Experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLOS Comput Biol*, *8*(7), e1002549. http://doi.org/10.1371/journal.pcbi.1002549

Jones, C., Hesterly, W. S., & Borgatti, S. P. (1997). A general theory of network governance: Exchange conditions and social mechanisms. *Academy of Management Review*, *22*(4), 911–945.

Kaye, J. (2011). From single biobanks to international networks: developing e-governance. *Human Genetics*, *130*(3), 377–382. http://doi.org/10.1007/s00439-011-1063-0

Kaye, J., Heeney, C., Hawkins, N., de Vries, J., & Boddington, P. (2009). Data sharing in genomics — re-shaping scientific practice. *Nature Reviews Genetics*, *10*(5), 331–335. http://doi.org/10.1038/nrg2573

Liebeskind, J. P., Oliver, A. L., Zucker, L. G., & Brewer, M. B. (1994). Social Networks, Learning, and Flexibility: Sourcing Scientific Knowledge in New Biotechnology Firms. *Institute for Social Science Research*. Retrieved from http://escholarship.org/uc/item/4480h6s7

Lucchi, N. (2013). Understanding genetic information as a commons: From bioprospecting to personalized medicine. *International Journal of the Commons*, *7*(2), 313–338.

Masum, H., Rao, A., Good, B. M., Todd, M. H., Edwards, A. M., Chan, L., … Bourne, P. E. (2013). Ten Simple Rules for Cultivating Open Science and Collaborative R&D. *PLoS Comput Biol*, *9*(9), e1003244. http://doi.org/10.1371/journal.pcbi.1003244

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the "tragedy of the commons." *Nature*, *415*(6870), 424–426. http://doi.org/10.1038/415424a

Mishra, A., & Bubela, T. (2014). Legal Agreements and the Governance of Research Commons: Lessons from Materials Sharing in Mouse Genomics. *Omics: A Journal of Integrative Biology*, *18*(4), 254–273.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review*, *23*(2), 242–266.

Nanda, S., & Kowalczuk, M. K. (2014). Unpublished genomic data–how to share? *BMC Genomics*, *15*, 5. http://doi.org/10.1186/1471-2164-15-5

Onwuekwe, C. B. (2007). Ideology of the Commons and Property Rights: Who Owns Plant Genetic Resources and the Associated Traditional Knowledge? In W. B. Peter & C. B. Onwuekwe (Eds.), *Accessing and Sharing the Benefits of the Genomics Revolution* (pp. 21–48). Springer Netherlands. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4020-5822-6_2

Ostrom, E. (2010). Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change*, *20*(4), 550–557. http://doi.org/10.1016/j.gloenvcha.2010.07.004

Overwalle, V., & Geertrui. (2013). *Governing Genomic Data: Plea for an "Open Commons"* (SSRN Scholarly Paper No. ID 2477897). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2477897

Parsons, M. A., Godøy, Ø., LeDrew, E., Bruin, T. F. de, Danis, B., Tomlinson, S., & Carlson, D. (2011). A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, *37*(6), 555–569. http://doi.org/10.1177/0165551511412705

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust. *The Academy of Management Review*, *23*(3), 393–404.

Six, B., Zimmeren, E. van, Popa, F., & Frison, C. (2015). Trust and social capital in the design and evolution of institutions for collective action. *International Journal of the Commons*, *9*(1), 151–176.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*, *6*(6), e21101. http://doi.org/10.1371/journal.pone.0021101

Tsai, W. (2002). Social Structure of "Coopetition" within a Multiunit Organization: Coordination, Competition, and Intraorganizational Knowledge Sharing. *Organization Science*, *13*(2), 179–190.

Zaheer, A., & Harris, J. D. (2005). *Interorganizational Trust* (SSRN Scholarly Paper No. ID 1256082). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=125608

Appendix 1. Open Science Grid

Project Information	
Name	Open Science Grid (OSG) www.opensciencegrid.org
Mission	Open Science Grid aims to support data-driven science by facilitating the redistribution of computational capacity across research institutions and communities. Computational capacity resources are pooled by the community, but managed and redistributed by the management team at OSG according to user conditions.
Field	Science
Brief history	While OSG foundational concepts began to take shape in 2002, the initiative was officially launched only in 2005. In the three-year gestation period, the infrastructure and the sharing mechanisms were conceptualized and defined.
Budget and funding source	OSG is jointly funded by the Department of Energy (DEA) and the National Science Foundation (NSF). The Department of Energy supports mainly small research projects, while NSF funds large projects. Combining funding from both is a mechanism to meet OSG diversified needs and balance institutional interests. Donor policies restricting private sector participation prevent OSG from directly engaging in private sector partnerships.
Size	The management staff includes around 30 members, supervised by an Executive Director.
Location	OSG has no physical site. Resources are located in the approximately 100 participating institutions. The management team is spread at various university institutions.
User community	Over 100 organizations are part of the OSG consortium, which includes thousands of users and operates approximately 800 million transfers per year. All actors involved are in the academic sector.
Technology	OSG does not own any computational resource, but it offers software and services to users, to access and utilize the common pool of computational resources that OSG members provide. The resources are easily accessible through a common grid that allows data to be transferred in different locations for processing. All OSG software tools are based on pre-existing open source tools that OSG tests or modifies in order to be productized. All developed tools are freely available to the members of the consortium.

Case Description	
Mission and main activities Goals Activities Boundaries	The aim of Open Science Grid is to re-allocate surplus of computational capacity among academic actors in different scientific fields to support data-driven and computational intensive projects. These might include analysis of large dataset or comparison of large amount of data across different datasets. For instance, OSG has been used in the study of genetic diversity and food security to understand "the performance of genetic clustering algorithms using simulated data".[11] OSG mechanism is simple: actors who have a surplus of computational capacity contribute to the pool, while actors who are in need of computational capacity can freely access and use the computational capacity in the pool. OSG acts as an intermediator of these transactions: researchers send their workload needs to OSG and OSG reallocates activities across the computational capacity available. The project is based on an "autonomy principle" according to which each actor can establish under which conditions to share or use resources. As intermediator, the OSG matches actors whose rules are compatible to allow the exchange of computation capacity. OSG does not own any of the resources that are shared within the project. The consortium members entirely provide them. OSG does not offer disc space. While computational capacity is relatively easy to share because it is never consumed, just temporarily occupied, disc space is more complicated to share because it is occupied for a long period of time. At the moment, OSG does not plan to modify its platform to share disc space.
History and drivers Foundation of the project Evolution Drivers Lessons learnt	The OSG consortium was officially inaugurated in 2005, but efforts to develop a grid able to support data-intensive research by re-distributing computational capacity across organizations date back to 1999. The need to redistribute capacity became clear when the amount of scientific data used for research grew beyond small institutions capacity and time-bound grants. For a group of scientists and informatics, resources sharing appeared as an efficient way to support data-intensive scientific research. The process to design the consortium was not straightforward. Despite the clear goals, the founders spent two years to design the consortium governance structure and sharing principles, before applying for funding. From a technical perspective, the main challenge for the design of OSG grid was the ability to serve research projects of different scales. At the same time, there was the challenge of creating a community around the project and engaging with actors in order to pool resources. The team decided to start by addressing the needs of small communities of actors with homogenous requests. Starting with a small group allowed identifying clear needs and achievable results, which stimulated motivation to participate. Demonstrating value was fundamental to engage actors in the project mission and attract new users. The OSG team collaborated with the small and homogeneous initial community in designing the platform.

[11] Source: http://www.opensciencegrid.org/genetic-diversity-and-food-security/

	However, the project was conceived to expand to diverse communities to be progressively integrated into the initial group. Gradual growth produced continuous buy-in and engagement. Communities are still important nowadays as they represent the basic unit to interact with the community.
Governance	
Macrostructure	
Membership Rules & benefits Private sector participation Developing country engagement	All actors participate in OSG on a voluntary basis. Most of them are academic scientists. The main motivation for them to participate is the belief that sharing resources will promote scientific innovation and discovery. There is a strong homogeneity of values within the consortium. Private companies are not OSG users because of privacy concerns and funding limitations. Private companies have applied some software tools developed by OSG for internal distributional systems. All OSG software tools are released with an open source license and companies can utilized them without restrictions.
Structure Governance bodies Management structure Teams and skills	The consortium is managed by a Council, which is composed of OSG lead stakeholders. The council is a "club": members of the council co-opt other members or remove them from the Council. Council members include all the OSG founders and representatives of selected organizations that joined subsequently. Members were co-opted to represent new disciplines. The Council includes IT professionals, software providers and scientists. The consortium elects the Executive Director. The Executive Director supervises around 30 employees, from diverse expertise domains, who are organized in four teams: (1) Executive; (2) Software; (3) Security; and (4) Operations. The community is structured around Virtual Organizations (VOs). VOs are defined by similar infrastructure needs, e.g. members of the same research project or the same university campus. VOs allow OSG to efficiently address community needs and reallocate resources. VOs represent an intermediate level of rules. Access to or use of a resource by a VO member is determined by both individual and VO rules. Being part of a VO is not a pre-requisite to access the OSG infrastructure.
Process	
Coordination Responsibilities Reporting & Monitoring	As OSG is a homogenous community, the management of the project does not require a complex, articulate structure to coordinate the initiative. Most of the coordination relies on the goodwill and informal relationships among actors involved.
Decision making	OSG does not have an advisory board. Strategic and scientific direction of the project are discussed and negotiated within the Council where the main stakeholders in the community are represented.

Decision making rules Autonomy	
Goal setting Strategic decision making	Goal setting mainly deals with defining new functionalities of the platform and IT development. OSG does not plan to enlarge its mission in the upcoming future.
Activities	
Outreach	For recruitment of diverse communities, the OSG team strategy centered on presence at major events and meetings in different fields of science and demonstration of scalability of the project. The approach is to connect with other projects rather than take them over.
Sharing Policies	
Sharing approach	OSG is based on the principles governing distributed IT systems. A distributed system is a grid that allows matching unused resources in one place with activities necessitating resources in another place. The project is based on an autonomy principle according to which: (1) actors can decide whether they want to use the capacity in the pool, contribute to the pool or both; (2) contributors and users are free to establish their own rules to contribute and use the shared resources. Contributors can decide who can access their resources, when, for how long, and in what quantity. Users can decide where and when to send data. By way of example, contributors can decide to donate their capacity only to small users. Users can choose to have their data processed only at a certain institutions, in certain hours. OSG holds no value judgement on rules established by contributor and users. The OSG system is designed to cater for all motivations and conditions to share. The OSG system supports this form of heterogeneity by applying an algorithm that matches users and contributors with compatible rules. Recognizing complementarity between users and contributors was fundamental in the design of the OSG platform. OSG has no monitoring system to verify who is contributing, who is using the resources and in what quantity. The assumption is that actors who do not want to share cannot be forced. It is the incremental effect that is deemed to foster the motivation to join the pool. OSG management has the fundamental role of guaranteeing that the system meets user requirements and preferences, and of solving technical problems.
Sharing rules	Sharing rules are established at the individual level.
Evaluation and Findings	

Course of the analysis	All interviewees agree on the value of the project in providing resources for data-intensive research. The project enjoys a strong reputation among users, who recognize effective and productive access to the pool. Users value the role by management team in setting up the collaboration. Interviewees generally confirm that the project does not intend to create a user community. Members have an online space where they can interact, share experience and feedback with the management team. However, members do not engage in further collaboration or discuss common interests and projects. Communities are exclusively functional to platform use. The management team shares a common understanding of the project and a strong conviction towards its vision and sharing approach. Autonomy is considered fundamental for sharing and so is OSG mission to accommodate user requests instead than trying to unify them.
Key findings	**Staged approach.** Projects start working with homogenous communities in order to avoid prolonged negotiations of goals and resources. The initial phase is complex, as needs and expectations are negotiated and incorporated in the infrastructure design. Once the project is established, it becomes easier to integrate heterogeneous communities and needs since: (1) there are already consolidated system and rules that prospective users can evaluate; (2) the management team is structured. **Autonomy.** Enforcing common rules for sharing might be counter-productive as individuals protect their own preferences and policies. Projects can leverage on complementarity between contributors and users needs to share resouces. **Project structure matches with shared resources.** OSG intermediates technical resources that do not require common standards and knowledge. As such, the project can operate with a simple institutional design centered on a technical management team and a representative Council with co-opted members.
Data Sources	
Interviewees	• Principal Investigator • Technical Director • Users
Attached documents	• OSG_Genetic resources project • OSG_Management plan • OSG_Organization • OSG_Presentation

Appendix 2. Structural Genomics Consortium

Project Information	
Name	Structural Genomics Consortium (SGC) www.thesgc.org
Mission	SGC is a non-profit, public-private partnership that aims to enhance pre-competitive research among pharmaceutical companies on cutting-edge but little investigated areas of human genetics research – such as protein structures and epigenetics. SGC supports and creates conditions for open collaboration between private sector and universities and promotes the free and open diffusion of research findings, which are made publicly available on the project website.
Field	Human health
Brief history	SGC received initial funding in 2004-2005[12] and at that time counted only one private company - GlaxoSmithKline - among its members. In its first phase (2005–2008), the initiative was focused on research on 3D structure of human proteins. In phase II (2009-2011), the consortium started working on chemical probes and epigenetics research. The new focus progressively attracted new partners. By 2011, around 10 companies were part of SGC. In phase III (2011–2015), the project maintained the same membership and kept exploring epigenetics research, while in phase IV (2015-ongoing) SGC is planning to enlarge its research focus to include clinical trials.
Budget and funding source	SGC is funded through membership fees. Each member provides an equal donation of US$7 to 8 million for a 4 to 5 year period. The amount of fees has changed over time. The amount is determined based on fees previously paid by other members or, more recently, based on grants received. The current fee amount is intended to match a grant received by the Innovation Medicine Initiative (Europe). The budget is entirely allocated on research projects and staff.
Size	SGC central structure includes 6 managers and coordinators, and around 40 PIs who coordinate SGC projects across its 4 different sites.

[12] Dates are approximated as interviewees reported them differently and no confirming evidence is available on SGC's website. All interviewees referred to the 4 Phases.

	In addition, SGC reports approximately 200 scientists, visiting scientists, and other support staff - including PhDs, post-docs and other researchers - who are funded by the project and work at one of its centers.[13]
Location	SGC central team is located at the University of Toronto, Canada. SGC's other campuses include the University of Oxford, UK, University of Campinas at Sao Paulo Research Foundation, Brazil, and University of North Carolina at Chapel Hill (SGC-UNC), USA. The University of Oxford and the University of Toronto are the two original sites. The sites in Sao Paulo and North Carolina have been added subsequently. Until 4 years ago SCG also had a lab in Sweden, which was closed due to lack of government funding.
User community	SGC membership includes 8 private pharmaceutical companies (Abbvie, Bayer HealthCare, Boehringer Ingelheim, Janssen, Merck, Novartis, Pfizer, Takeda) and several public and non-profit institutions that also contribute as members (Canada Foundation for Innovation, Sao Paulo Research Foundation, Genome Canada, Ontario Ministry for Research and Innovation, the Wellcome Trust).
Technology	SGC does not provide any specific technology to its members.
Case description	
Process	
Mission and main activities Goals Activities Boundaries	SCG aims to accelerate scientific research in human genetics for health. It does so by promoting open, collaborative and pre-competitive research among pharmaceutical companies and universities in the field of human genetics. SGC acts as a trusted broker to facilitate private company and university collaboration. It reduces concerns – IP, privacy, mistrust – that generally characterized those collaborations by defining clear rules and by intermediating relationships among actors involved. SGC leverages on private sector's concerns that property right-based R&D models are not anymore suitable for advanced genetics research. In most of cases, companies sustain heavy research investments for activities or products that do not produce profits. The simple idea of SGC was to pool together financial resources to support research in less competitive areas – i.e. compounds for target validation – that might be beneficial for all companies without eroding their strategic advantage. The ratio risk-investment of collaborative projects in SGC is lower than the one of companies' internal programs, as long as companies are able to maintain their competitive advantage. SGC role is to support companies towards cutting-edge, research through collaboration while guaranteeing that no proprietary information is

[13] www.thesgc.org/about/mini_faq

	revealed to other companies. Companies avoid investment duplications, benefit from a wider research network and expand their areas of research. SGC involves university partners that collaborate with industry on R&D topics. By participating to SGC, universities are able to access significant funding for research on cutting-edge science topics and build connection with industry. SGC engages with both industry and university to promote an open approach to sciences, based on the free publication of research findings. All results from SGC research activities are freely available online without use restriction.
History and drivers Foundation of the project Evolution Drivers Lessons learnt	SGC history is divided into four phases of 3 to 5 years each. SGC originated from the realization that human biology science requires multi-organization partnerships. The initiative was initially funded by public organizations, such as the Wellcome Trust, Genome Canada, the Ontario Ministry of Research and Innovation and the Canada Institute for Health Research (CIHR). Except for CIHR, all other organizations are still members of SGC. CIHR resigned from membership two years ago, but still funds some projects. SGC initially focused on sequencing the 3D structure of human and parasite proteins of interest for human diseases. From this activity comes the name of the project, *structural genomics* consortium. After 2 years of activity, the only private company involved in SGC at the time, GlaxoSmithKline (GSK), proposed to launch a new project to focus on chemical probes for target validation in the pharmaceutical industry. SGC accepted the project as it allowed building on prior knowledge to explore a new research area of potentially high scientific value, i.e. the epigenetics. The project, which involved also the Ontario government, was a success and led to SGC second phase. Phase II opened with renewed efforts on epigenetics research that attracted new private companies to join SGC. At the beginning of phase III in 2011, the SGC had around 10 private company members and included a new project to develop renewable antibodies for epigenetics proteins in collaboration with life technologies companies. Phase III represented the consolidation of SGC activities and a progressively shift towards clinical oriented research, which is at the core of the current Phase IV. The evolution of SGC goals has been driven by a number of factors. First, SGC has brokered industry needs and developed projects in line with pharmaceutical companies' research direction. Second, SGC has successfully identified cutting-edge but under-investigated research topics. The majority of those topics are too complex for a single organization to investigate comprehensively and to produce return on investment. By focusing on those topics, SGC has managed industry competitive concerns by avoiding interference with companies' core R&D activities. Finally, SGC has been able to leverage on accumulated skills and competences by progressively enlarging its research areas. The switch towards epigenetics research is an example of the interplay among those factors. The new project was launched under suggestion of a private partner which recognized in epigenetics a high potential but uncertain research field. The consortium accepted

	the topic as skills already developed within SGC laboratories were suitable to the new endeavor. For the growth of the consortium, it was also key to gain a leadership position vis-à-vis private companies through norm setting. A fundamental factor was to start with a single partner. By working with a single partner, SGC was able to negotiate its own rules (i.e. open access to research findings) without engaging in time consuming discussions among a large set of actors with divergent interests. When other members decided to join, they had less ability to influence the existing rules. This process reduced tensions and allowed the initiative to quickly grow and continue producing research outputs. Finding common ground among all partners leveled expectations and set reasonable goals. Universities and companies might still have different approaches to research and research methods but setting the rules transparently and discussing reciprocal expectations openly avoided misunderstanding.
Governance	
Macrostructure	
Membership: Rules & benefits Private sector participation Developing country engagement	Membership includes both private and public organizations. To become a member, companies, non-profit or public organizations are required to pay a membership fee. Membership gives the right to nominate a representative to the Board of Directors, to nominate a member within the Scientific Committees and to have on-site scientists within SGC centers. Although there is no established maximum number of members, the high membership fee prevents smaller organizations from joining SGC. As a consequence, scientific capacity within SGC members is relatively uniform. Small companies have tried to join the consortium by proposing a different type of collaboration that would not entail membership fee. The SGC has decided not to accept the proposals in order not to alter the fee-based membership mechanism that is believed to ensure equality among all members. SGC generally stipulates individual agreements with companies as multi-parties agreements are considered too complex to design. Although agreements are generally standardized, a degree of flexibility in negotiating with companies is applied – i.e. by adapting agreements to company practices and wording. . Private companies draw several benefits from joining SGC as members. The projects ensure access to high quality research findings that companies can directly input into their own projects. To maximize usability, SGC requires universities to apply industry protocols and standards to SGC research activities. All research within SGC is designed to be reproducible in other laboratories. Member companies have access to results with a 24-month embargo before publication. In addition, SGC projects increase the reputation of private companies among universities and offer companies the opportunity to access research networks. With SGC, Companies are connected with research outside of their labs and kept abreast of scientific developments through talks, papers and publications.

	Research institutions draw benefits too. Academic scientists are exposed to different research approaches and new project opportunities. Researchers maintain the ability to publish research out of SGC projects and, in the light of the time that peer review processes require, are not negatively affected by the embargo period.
Structure Governance bodies Management structure Teams and skills	*Strategic and advisory boards* SGC is governed by a Boards of Directors that includes a representative of each member of the Consortium, plus the Chief Scientists of the four SGC centers. The Board decides SGC activities on a yearly basis. The work of the Board is supported by two Scientific Committees (the Protein Structure Scientific Committee and the Epigenetics Chemical Probes Scientific Committee). The Scientific Committees provide strategic advice and direction on SGC research activities, and oversee research quality and reliability. For instance, the Scientific Committees approve all chemical probes developed by SGC prior to their online release. Members of the two Scientific Committees are representatives of SGC member organizations involved in the respective projects. The External Scientific Advisory Board is tasked with reviewing the overall quality of SGC research before release. The External Board includes worldwide recognized experts in multiple fields. *SGC internal structure* A CEO leads the management structure, and is responsible for the coordination of all SGC centers and projects. Each SGC center is directed by a Chief Scientist who coordinates all research teams within the center. Each team works on a different project and is managed by a PI. Scientists within the labs work closely with scientists in private companies, including by reciprocal hosting. All scientists working for SGC are faculty members at their university, but are entirely paid by the consortium. *Project management* SGC has a project manager that coordinates all projects with private companies. The project manager meets with each company representative every 6 to 8 weeks. Since meetings are private, companies can disclose information that they do not want to reveal to other companies but are useful for SGC to coordinate activities at the consortium level. In some cases, the Chief Scientist involved in the project is invited to participate. The Joint Management Committee (JMC) ensures coordination and communication across projects. One representative for each company is invited to attend. The SGC project manager reports on progress by projects in aggregate mode, to protect companies' data and information. The JMC is an important vehicle of information to members and of continuous confidence by members in the validity the scientific data.
Processes	

Coordination Responsibilities Reporting & Monitoring	Coordination is a key function of SGC since the consortium has to ensure that private companies' secrecy requirements are matched and research is carried according to high-quality standards. Companies are assigned to projects by SGC, which ensures that no more than two companies work on the same chemical probe. Companies are aware of all activities within the consortium but do not know which company is working on which activities. Only SGC management staff has access to such information. In Board meetings, information on projects is presented in aggregate or anonymized forms. This protects companies' competitive information. SGC centers have to report quarterly to the Board of Directors on projects status. The project manager coordinates the relationships with external partners and SGC sites. Disagreement and various issues related to implementation within projects are resolved by iteration between the project manager and companies first, and with the JMC and the Board in second and third instances. Focal points within each company coordinate activities between the company and the SGC. They make sure that the value of the collaboration is well understood by internal management and regularly communicate benefits and results of SGC activities.
Decision making Decision making rules Autonomy	All consortium activities are negotiated within and decided by the Board of Directors. The SGC Director oversees SGC research activities and executes decisions taken by the Board. Chief Scientists supervise and decide scientific activities within their Centers.
Goal setting Strategic decision making	Consortium-level goals are established on a 5-year basis by the Board of Directors. Each company can propose a compound (protein or chemical probe) of interest. Proposals are not disclosed to other companies. SGC management selects target compounds and allocates projects to companies by matching proposals as much as possible. Goals are reviewed by the Board on an annual basis based on available resources.
Activities	
Outreach	SGC has an Ambassador Program for SGC researchers (generally PIs) to give talks and meet with scientists at private companies. This allows companies to familiarize with SGC activities before deciding on membership. Scientist-to-scientists exchanges of tacit knowledge within SGC formalized structure are highly valued by companies and universities. Some companies require reports about visits done to company sites to showcase the value and the intensity of the collaboration with SGC to the top management.

Sharing Policies	
Sharing approach	SGC distinguishes between internal sharing of company information and external sharing of information produced by SGC activities. On the one hand, SGC protects companies' information that are neither shared among SGC members nor externally. Companies are free to decide to which extent they want to share their information in order to protect their competitive advantage. On the other, all results produced by SGC research activities are published online without use restrictions.
Sharing rules	*Sharing research results* Shared findings include, among others, protein structures, chemical proves, and antibodies. Experimental protocols are also shared so that other researchers and scientists can reproduce SGC findings in other contexts. Patentability – including by SCG members – of the above research outputs is excluded by agreement. This rule is non-negotiable by members and applies private and public members. The non-negotiable nature of the rule helps its implementation as SGC does not give members opportunities to discuss exceptions or other arrangements. Companies are granted a 24-month embargo period to take advantage of SGC research outcomes (the Board has extended the original 12-month period to 24 months). To be released, findings have to meet precise selectivity and potency criteria. Results are accurately checked and validated by the Scientific Committee and the External Advisory Board. Only members have access to pre-released results. By not imposing use restrictions, SGC aims to to accelerate scientific discovery and drugs development in the field on human health. Upon request, SGC makes available constructs, DNA materials, samples of the chemical probes, and experiment details. The quality of data and the transparency of the initiative have gained SGC a strong reputation among the scientific community. *Sharing companies information* Sharing on issues other than SGC research findings is very limited and controlled. Companies do not share their internal data – i.e. chemical compound structures - and when they do, SGC guarantees that only essential information is shared. For instance, company data libraries are blindly screened to select only a small number of structures and compounds of direct relevance to the research objective. Only selected structures and compounds are shown to academic scientists. Libraries may be blindly compared to identify common research areas. In meetings, data are presented only in aggregate and codified form so that no proprietary information is shared.
Evaluation and Findings	
Course of the analysis	All interviewees consistently report SGC activities and goals and agree on the overall effectiveness of the project. SGC meets the needs and expectations of partners and other actors involved by creating a truly collaborative environment.

	All interviewees also highlight the importance of clearly defined rules by SGC, to build trust across the consortium and to engage actors into reciprocal, collaborative behaviors.
Key findings	**Engaging with diverse sectors: rules.** SGC has been able to engage with actors from different sectors, namely pharmaceutical companies and public universities. Clearly defined rules and leveled expectations have been the most important enabling factors. Actors are able to engage on a peer-to-peer basis as rules allow them to focus attention on scientific issues rather than other more sensitive ones, such as property aspects or goal settings. SGC acts as a mediator by facilitating and regulating research activities and ensuring that key information is protected. **Engaging with diverse sectors: common ground and value.** SGC has promoted a shared understanding of research quality and standards, which is feasible by universities and meets company requirements. Shared understanding allows for ensuing collaboration to create values for all parties involved. **Engaging with diverse sectors: homogeneity.** While SGC has engaged actors from different sectors, there is a certain homogeneity among them in terms of capacity. The membership fee drove such homogeneity and as a result SGC does not have to deal with capacity building. **Sharing with private companies.** Sharing data and information with private companies is challenging. Nevertheless, SGC shows that it is possible to work on common projects with private companies and set rules for making research findings open and freely available. Companies agree on financing research and sharing common findings if: (1) there is an embargo period to maintain competitive advantage; (2) information is adequately protected; and (3) no interference with company privileged research areas exists. **Showcasing value.** SGC representatives constantly show the return on investment, in terms of new products and collaboration with university. A liaison contact within each company facilitates such communications. Public events (talks, presentations) maximize reputational benefits.
Data Sources	
Interviewees	• SGC Director • Management staff • Chief Scientist at SGC centers • SGC members
Attached references	• SGC_Governance
References	Masum, H., Rao, A., Good, B. M., Todd, M. H., Edwards, A. M., Chan, L., … Bourne, P. E. (2013). Ten Simple Rules for Cultivating Open Science and

	Collaborative R&D. *PLoS Comput Biol*, *9*(9), e1003244. http://doi.org/10.1371/journal.pcbi.1003244 Williamson, A. R. (2000). Creating a structural genomics consortium. *Nature Structural Biology*, *7 Suppl*, 953. http://doi.org/10.1038/80726

Appendix 3. iPlant

Project Information	
Name	iPlant Collaborative (iPlant) www.iplantcollaborative.org
Mission	iPlant is a downloadable, open source data management platform which provides biology scientists with informatics tools for the management, analysis, sharing, visualization and storage of large amount of genetics data.
Field	Food and agriculture
Brief history	iPlant was established in 2008 thanks to a NSF 5-year grant to: (1) facilitate access to advanced IT tools for biology scientists; and (2) enhance collaboration among scientists on data-intensive research projects. The grant was renewed in 2013 for additional 5 years.[14] At the beginning of 2016, iPlant was re-branded in CyVerse in order to "to emphasize its expanded mission" towards all life sciences.[15]
Budget and funding source	The NSF granted iPlant US$ 100,000,000 distributed over 10 years, up to 2018. The future financial sustainability of the project is currently discussed within iPlant and with NSF. iPlant is considering several market-based options including developing a fee-based system for use of the platform, offering consulting services. On the other, iPlant is trying to diversify and enlarge its services to other scientific fields, partnering with external organizations to obtain other grant funding, and negotiating further funding with NSF to continue providing researchers with cyber-infrastructure.
Size	The project is led by a team of 7 Co-PIs and a 9-member Scientific Advisory Board. The project includes around 100 collaborators, among which scientific analysists, project coordinators, researchers, engineers, communication and outreach staff, student and faculty members.[16]
Location	iPlant is a US-based project. Teams are located at seven institutions: Cold Spring Harbor Laboratory, University of Arizona, University of Texas - Austin, Texas Advanced Computing Center, University of California – Santa Barbara, University of North Carolina – Wilmington, NCEAS – UC Santa Barbara.

[14] NSF grants No. DBI-0735191 and DBI-1265383
[15] Source: http://www.cyverse.org/about.
[16] Retrieve from www.iplantcollaborative.org/about-iplant/the-project/people on January, 2016

User community	The iPlant platform is used all over the world. Most of users are in the UK and the US, but a significant part of them is also located in Africa and Asia, including China, India, Malaysia, Korea and Egypt.[17] iPlant users are academic and public research institutions. While at the beginning the platform was targeting only biology scientists, now iPlant offers products and services to researchers in several life sciences, including both plants and animals genetics. No private actors are currently using the platform.
Technology	iPlant offers an online, free and open source platform for the management, analysis and sharing of data. iPlant users can upload their data online and utilize iPlant tools for analysis, data management and visualization. iPlant does not offer storage space. Data can be uploaded online in iPlant storage space only if they are used for analysis in the iPlant environment. At the current stage, the platform manages multiple data categories. The integration of new data categories requires intensive, but not complicated, work for the core infrastructure team. The brand "Powered by iPlant" identify projects that leverage on iPlant infrastructure – i.e. iPlant tools and platforms – to provide their users with services.
Case Description	
Mission and main activities Goals Activities Boundaries	iPlant focuses on three main goals: (1) providing scientists with an adequate cyberinfrastructure for data intensive life science research; (2) supporting collaboration among scientists; and (3) encouraging data sharing. iPlant is defined by the products and services it offers to users. Products are the iPlant infrastructure and the computational tools that allow scientists to manage, analyze, and storage their data. Services include technical and scientific support by iPlant staff; brokerage activities to connect users with other scientists or technicians for collaboration and support; partnering with iPlant users for the development of grant proposals; and specialized support for standards and metadata within scientific communities. While iPlant activities are mainly shaped by community's requirements and needs, the Executive Team has decided not to engage in data analysis or production of research findings, in order to focus on iPlant core products and services.
History and drivers Foundation of the project Evolution	The project was launched by a small group of bioinformatics and scientists to advance computational tools for data management in plant genetics. Given the expertise of its members, the initial team was able to link the scientific vision of the project with the technical vision, designing the higher architectural concept of iPlant and defining how it would relate with its community. Nevertheless, it was fundamental for the success of the initiative to progressively enlarge the team with new expertise.

[17] Not exhaustive list.

Drivers Lessons learnt	The setting up of the project team was a fundamental step to launch iPlant. The initial team members selected researchers with a widely recognized reputation in their field. Selecting a team on the basis of individual skills was fundamental to avoid interference by individual interests with project goals and to build a trustworthy relationship with the community. In addition, new experts were added to the team only upon request and consensus by other team members, in order to ensure group cohesiveness, trust and shared vision. Those are essential assets for iPlant as team members work in a geographically dispersed environment and with a high degree of autonomy. Both managerial and technical competences were injected into the team, to set project milestones and handle pressure for delivery, and to expand through education, training and outreach (for this latter goal, NSF suggested iPlant to partner with Cold Spring Harbor Laboratory). In its initial phase, iPlant had to persuade scientists of the value of the platform. At that time, biology was less data-intensive and fewer scientists were seeking data management tools. In addition, there were technical constraints to overcome to generate a user-friendly, intuitive platform. From its second year, the iPlant team organized a series of workshops and conferences to gather experts and scientists together, and involve them in the design of the platform. Through this in-take mechanism, while biology was becoming increasingly data-driven, the iPlant team was able to collect feedback and ideas, and translate scientist needs into user-friendly IT tools. Along with the setting up of the team and collecting feedback from the community, the definition of iPlant boundaries was a crucial step to safeguard its implementation and effectiveness. iPlant was exposed to several pressures from potential partners and collaborators. It was increasingly impossible to respond to all of them and design a feasible working plan. The iPlant team clearly set the primary mission, i.e. to enable science and not to do research. iPlant refused engaging in data analysis and data production, and did not collaborate with private sector, whose goals and working schedule were not aligned with its owns. A clear definition of the project mission and boundaries remains important in the current phase. In 2014-2015, the project has significantly augmented its technological capacity and now provides a comprehensive platform for data management, which has progressively attracted researchers from new fields of science. The Executive Team in collaboration with NSF has decided to re-branded the project as “CyVerse”. The change symbolizes the universality of scientific data that the platform can manage. In addition, iPlant is trying to transition towards a financially sustainable, market-based model. Maintaining the focus on core goals is essential to position the platform in the market and articulate clear value propositions to customers.
Governance	
Macrostructure	

Membership: Rules & benefits Private sector participation Developing country engagement	iPlant does not have a formal membership system. Joining the iPlant community is free and members can use the platform upon registration on iPlant's website. Registered users have access to the iPlant cyberinfrastructure and the data that are publicly available in iPlant's database. There is not any governance body or mechanisms for representation of users in iPlant decision-making processes. Users are mostly from government, academia and non-profit organizations. iPlant has made attempts to cater for private sector but with no success. Security, project scale, time scale, reproducibility are typical private sector needs that are difficult to match with iPlant vision. Small companies are more flexible and willing to negotiate common goals but collaboration remains difficult. Developing countries are involved as users of the platform. iPlant offers a set of support activities – in-site training and workshops – that are dedicated to developing country scientists.
Structure Governance bodies Management structure Teams and skills	An Executive Team leads the project, including a PI, four Co-PIs, distributed across all sites of the project, and two directors – of public-private partnership and of the DNA Learning Center. The PI is responsible for the strategic direction of the project, while the Co-PIs are responsible for cyber-infrastructure development and scientific engagement activities on their sites. A Scientific Advisory Board advises the Executive Team about the strategic direction of the initiative and the allocation of resources among different projects. The Executive Team manages the operational staff, which is responsible for delivery. All other collaborators are organized in a matrix structure, according to two dimensions: the site where they are located and the skill-based team in which they work. Teams are designed around three main skills: Cyberinfrastructure Development (CD), Scientific Engagement (SE) and Education, Outreach and Training (EOT). Each team includes at least one member from each site, who coordinates other staff members in its location. The teams are the following: (1) The core services team (CD), responsible for daily maintenance and functioning of the platform (help desk and support system); (2) The core software team (CD), responsible for development and maintenance of APIs, tools and software; (3) The APIs team (CD), responsible for the iPlant computing center and the federation with other projects; (4) The science analyst team (SE), responsible for the interface between developers and the scientific community - the team is composed by PhDs conversant with both research and informatics who connect with the scientific community, e.g. by travelling to conferences, organizing working groups, collecting inputs from the community about what iPlant should or should not do, and supporting scientists in utilizing iPlant tools.

	(5) The education, training team, outreach (EOT), responsible for coordinating EOT efforts across all sites, reaching out intermediate users, and providing training to new users, including graduate and undergraduate students. iPlant has supported the creation of satellite projects to deliver iPlant services (computational and analysis capacity) at the local level. Those projects are completely autonomous. The advantage of local projects is their ability to provide additional computational and analysis capacity to their local users. Currently, there is one local project in UK and there are negotiations to open another local iPlant project in Chile. iPlant has collaborated with external projects for the development of specialized IT platforms for the management and sharing of science data. Those projects are labeled under the name "Powered by iPlant" and are completely autonomous from the iPlant management structure.
Processes	
Coordination Responsibilities Reporting & Monitoring	Internal coordination is ensured through regular meetings among members of the Executive Team, among members of each team (across sites) and within each site. The Executive Team generally meets every week, and teams have to report to executive members at least quarterly. Team-based meetings are strongly encouraged to facilitate coordination across all geographical locations of the project. Meeting are generally self-organized by members according to their needs. The operational staff updates internal monitoring documents and track progress by each team. Coordination with external partners is by the Executive Team, which is responsible for partnerships. The Executive team also guarantees coordination with partner projects, i.e. the "Powered by iPlant" projects.
Decision making Decision making rules Autonomy	The Executive Team is responsible for strategic decision-making – i.e. on partnerships – and has the final authority on all decisions taken within the project. Decisions are taken collectively. In the decision-making process, the Executive Team closely collaborates with the NSF Plant Science Cyberinfrastructure Collaborative (PSCIC) and the NSF iPlant Program Officer. It also receives community input and strategic advice on project direction from the Scientific Advisory Board. An external evaluator, East Main Evaluation & Consulting, LLC (EMEC), supports the Executive TEAM.[18] Teams make decisions on the daily management of their area of competence.
Goal setting Strategic decision making	Goals are established in the grant proposals that iPlant submitted to NSF in 2007 and 2013. To cater for community needs and project growth, the Executive Team frequently interacts with its community and the NSF to revise iPlant goals.

[18] http://www.iplantcollaborative.org/about-iplant/the-organization/leadership

Activities	
Technology management	iPlant has been developed by assembling already existing software from other NSF-funded projects or DEO projects. Most of new tools are not entirely developed by iPlant but are created in collaboration with the community. Thus, the management of the platform requires continuous collaboration with external partners. Moreover, as iPlant aims to respond to community needs, a large part of the platform management includes collecting feedback from users. The IT team carries technology assessments to evaluate and rate, and then prototype and integrate tools.
Sharing policies	
Data sharing approach	iPlant goal is to encourage data sharing, rather than enforcing it. iPlant focuses on enabling factors, such as metadata and standards. Standards are developed by iPlant for generic data while, for domain-specific data, iPlant is encouraging a community-driven process where researchers have to design their own standards and metadata. iPlant is willing to collaborate with them and provide technical staff support during the process, but avoids direct involvement in specific projects as it would require too many resources (i.e. human and time resources) for an activity that is judged outside the boundary of the initiative. iPlant enables data-driven science through technological tools but does not engage in producing scientific content. Moreover, the Executive team values direct engagement by scientific communities in the design of standards. Engagement facilitates adoption, as standards are defined bottom-up and not top-down. iPlant offers a facilitation platform and ontology specialists to support scientific communities which, however, decide autonomously. The design of common standards and metadata will increase its importance for iPlant as the project moves forward. For the future, iPlant aims to encourage researchers to deposit their data into its Data Commons platform. The underlying philosophy is that open data are useful only if they are able to create value for someone else. Hence, the platform will not make available all iPlant data. It will enable researchers to share the data that might have value for other researchers. The project will not enforce any rule but will apply NSF requirements regarding sharing of publicly funded data.
Data sharing policies	As iPlant does not want to enforce data sharing rules that might prevent users from utilizing the platform, users are free choose whether they want to keep their data private, share with a group or make them public.
Other resources sharing Infrastructure	The use of iPlant resources is free. NSF required that all resources be publicly and globally available. iPlant set some basic rules to regulate platform access and availability of computational resources to avoid resource misuses and consumption (i.e. CPU or storage space). At the beginning of the project, iPlant experienced the risk of exhausting its computational capacity due to overconsumption in Eastern Asian countries. Multiple users were connected at the same time from the same

	location. iPlant requires all users to be registered with an institutional email address and students can access the platform only after their mentor has sent a letter to state how and for what purpose iPlant resources will be used. iPlant does not allow users to store their data for long periods of time, unless they are actively used for analysis.
Evaluation and Findings	
Course of the analysis	Interviewees share a clear understanding of iPlant mission and goals. They often highlight how coherence towards these goals has been fundamental throughout the project to maintain a unified community and its commitment towards the initiative. All interviewees also report that boundary definition was critical. Until activities and resources remained undefined, it was difficult to establish a strong leadership team and the initiative was subject to several pressures from external stakeholders. Definition of goals and coherence towards them were key elements to stabilize the project. Among interviewees, there is also a clear understanding of iPlant's data sharing approach. All of them recognize user autonomy in data sharing decisions and understand that their task is to facilitate the process, rather than monitor or control it.
Key findings	**First, enabling and second, sharing.** iPlant enables data sharing among scientists by providing an adequate IT infrastructure and collaborating with different communities to create metadata and standards. iPlant believes that it will be able to create – in the long run – a community that recognizes iPlant as a reliable depository of open and freely available data. **Demand-driven.** For a project to be accepted by the scientific community, it is important to engage with potential users since its inception. **Autonomy**. iPlant refuses a top-down approach to metadata and standards development. It offers support to scientific communities to develop their own metadata and standards.
Data Sources	
Interviewees	• iPlant PI • Co-PIs • Project coordinators
Attached documents	• iPlant_Original NSF grant proposal • iPlant_Advisory Board • iPlant_Team • iPlant_Quarterly goals
References	Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., … Stanzione, D. (2011). The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Frontiers in Plant Science*, *2*. http://doi.org/10.3389/fpls.2011.00034

Appendix 4. Global Alliance for Genomics and Health

Project Information	
Name	Global Alliance for Genomics and Health (GA4GH) www.genomicsandhealth.org
Mission	The Global Alliance for Genomics Health aims to accelerate research in genomics for human health by promoting data sharing and data accessibility. The Global Alliance works to "establish, broadly disseminate, and advocate for the use of interoperable technical standards for managing and sharing genomic and clinical data".[19] GA4GH also supports and develops projects that demonstrate the value of increasing data sharing and data accessibility.
Field	Human Health
Brief history	January 2013 – The idea of GA4GH is proposed during a meeting by 50 experts from 8 different countries. Their goal was to tackle some of current challenges in genomics research, in particular the lack of harmonized approaches for sharing genomic and clinical data in an "effective, responsible and interpretable manner"[20]. January to June 2013 – A first draft or a White Paper describing the goals and mission of the Alliance is circulated among possible members, along with a non-binding Letter of Intent (LoI). June 2013 – The Alliance is officially launched with the subscription of the LoI by 70 organizations. March 2014 – First face-to-face plenary meeting is held at the Wellcome Trust in London. The Constitution of GA4GH is approved.
Budget and funding source	The Alliance is mostly funded by three institutions: the Ontario Institute for Cancer Research, Canada, the University of Cambridge, UK, and the Broad Institute, Massachusetts, USA. Other organizations provide smaller funding – including the Wellcome Trust, the Sanger Institute and Genome Canada. The Alliance prefers to receive small grants to maintain its flexibility and not to compete with its own members for bigger funding.
Size	The organization reports 386 members.[21] It is led by a Steering Committee of 16 members, the GA4GH Director and a team of 8 managers. No information is available on management additional staff members.

[19] Source: www.genomicsandhealth.org/about-global-alliance
[20] Source: www.genomicsandhealth.org/about-global-alliance/history
[21] Source: www.genomicsandhealth.org/about-global-alliance

Location	GA4GH is hosted at the three main funding institutions.
User community	GA4GH reports organization and individual members in 38 countries, including nonprofit, private and public organizations, in multiple fields, from bioinformatics to pharmaceutical companies and publicly funded research projects.
Technology	GA4GH does not focus on the development of a specific technology. Nevertheless, some of its demonstration projects include the development of technological tools. The Beacon Project provides a query that might be integrated into genomics databases to facilitate users' search for data. The Matchmaker Exchange project offers a platform that facilitates the matching of cases with similar phenotypic and genotypic profiles across multiple datasets through a standardized application programming interface (API).
Case Description	
Mission and main activities Goals Activities Boundaries	GA4GH aims to gather a variety of actors operating in the genomics health field to tackle challenges that are affecting the growth of data-driven genetics research. The Alliance mainly focuses on: (1) promoting accessibility and integration of genomics data across different cyber-infrastructures; (2) designing common policies and standards for data sharing; and (3) diffusing best practices and ideas across diverse communities in the field of genomics for human health. The work of the Alliance is organized around community needs and inputs. Community needs are identified by the Steering Committee and addressed through the activities of the Working Groups and the demonstration projects. Working Groups are thematic forums that discuss critical aspects of sharing and access to data, from regulation to ethical and security issues. The demonstration projects leverage on the tools and solutions proposed by the Working Groups to develop and realize concrete projects that showcase the value of the GA4GH principles. In general, the focus of the Alliance is more on interoperability issues that affect data sharing, rather than on the creation of data exchange relationships. The Global Alliance "does not itself generate, store, or analyze data, perform research, care for patients, or interpret genomes".[22] In the future, the Alliance would like to create a pre-competitive space to allow private actors to collaborate on projects.
History and drivers Foundation of the project	The Alliance was designed by a group of 50 experts in January 2013 with the aim of discussing challenges and opportunities in the genetics research for human health. Among several issues, a general consensus emerged concerning the need of designing harmonized approaches to effectively share data among actors in the field. The initial target was to gather 25 organizations around this mission.

[22] Source: www.genomicsandhealth.org/about-the-global-alliance/frequently-asked-questions#t355n6319

Evolution of goals Drivers Lessons learnt	In the pre-launch phase (January - June 2013), the initial group was able to gather together 73 organizations. Most of the energy and the enthusiasm in this first phase came from individuals and organizations which had interests in line with the Alliance purposes. All initial members were non-profit organizations (mainly, genomics research institutions). The only requirement for them was to sign a non-binding LoI to collaborate. The initial group started its collaboration by writing a White Paper that describes the objectives and activities of the Alliance. Producing the White Paper was an opportunity to discuss the perspectives of different partners involved and to negotiate the future direction of the Alliance. It consolidated the first group of partners. The White Paper iterations were instrumental to receiving feedback and suggestions from the community and harmonizing expectations. It also helped management refine the direction and focus of the Alliance. After the official announcement of the Alliance in June 2013, the membership grew rapidly and the first partner meeting in 2014 was attended by over 200 organizations and individuals in the field to take stock of achievements made by the Alliance and discuss critical working areas. The growth of the Alliance is not a causal process. Rather, the founders have driven the Alliance to progressively attract diverse groups. The founders decided to gradually increase the heterogeneity of actors and minimize tensions and conflicts through such a gradual expansion. The first expansion of members was oriented towards increasing the geographic diversity of GA4GH members. The Global Alliance supported efforts to expand represented nationalities from UK, USA, Canada and Australia to Japan and Europe. The focus on this phase was still on developed countries. In the second phase, the Global Alliance became more concerned about knowledge diversity. The Alliance targeted not only research institutions, but also clinical institutions and private companies. In the most recent phase, GA4GH is targeting developing countries (i.e. African research institutions). A key achievement and further driver of the Alliance was the Framework for Responsible Sharing of Genomics and Health-Related Data published in September 2014. The document attracted further partners and consolidated the harmonized approach towards data sharing and safeguarding of fundamental individual rights. The Framework is intended to be a reference document for the entire field.
Governance	
Macrostructure	
Membership: Rules & benefits Private sector participation	Membership is open to both individuals and organizations active in the field. As GA4GH wants to maintain an inclusive approach, there are no fees associated to membership, nor obligations. Actors are free to decide their level of engagement within the initiative, from active participation in the Working Groups to simple subscription to a newsletter. GA4GH members can be elected to the Steering Committee and can propose members to be elected. However, members are generally more interested in joining the Working Groups and the demonstration

Developing country engagement	projects. Each member is free to join one or more groups or projects and to determine the level and modality of contribution. Membership is highly heterogeneous, from research projects to universities and non-profit institutions and private companies. Some companies are willing to collaborate on the development of tools for data sharing, accessibility and interoperability (e.g. through APIs).
Structure Governance bodies Management structure Teams and skills	The Steering Committee, composed by 16 voluntary members is the decision-making body of the Alliance. Steering Committee members are voluntary experts in human genomics or clinical research. They are nominated by GA4GH members and voted by other Steering Committee members. The Steering Committee appoints its Chair and the Executive Director of GA4GH. The Chair of the Steering Committee and the Chairs or Co-Chairs of the Working Groups form the Executive Committee. The Strategic Advisory Board, which has recently been established, is led by an independent chair (a senior scientist who is widely recognized in the field) and is composed by the Directors of the funding agencies, the Directors of the host institutions and a selected number of directors from other institutions. It is tasked with strategic oversight on the development of the Alliance. The Executive Director leads a Secretariat composed of 4 Working Group managers/coordinators, one membership coordinator, one communication lead and one administrative assistant. Each manager/ccordinator serves the Secretariat from its home institution. The Working Groups are established by the Steering Committee with the advice of Directors, and managed/coordinated by one of the GA4GH managers and one representative of the Steering Committee. There are 4 Working Groups in place: 1. Clinical – it focuses on clinical data sharing and integration of clinical data with human genomics data. It collects existing best practices and develops ontologies for curating and sharing of clinical data. 2. Data – it focuses on infrastructure, including data representation, storage, and analysis tools. It also designs interoperability standards. 3. Regulation and Ethics – it develops harmonized approaches to privacy, consent, and policies and data sharing agreement templates in line with the 2014 Framework for Responsible Data Sharing. 4. Security – it focuses on security, user access and audit function of shared data and data repositories. It develops minimum standards and guidelines for data security and protection. Within working groups, experts and GA4GH members cooperate for the development of standards and tools that GA4GH makes freely and publicly available to the whole community. Participation in the Working Groups is open and voluntary. The size of the Working Groups varies, up to over 100 members, although not all of them are active. Working Groups are further organized in small task teams.

	The Alliance manages three demonstration projects that implement tools and policies developed in the Working Groups to show their value and develop in-take mechanisms (see below the section on *project implementation and design*). The projects have autonomous governance structures.
Processes	
Coordination Responsibilities Reporting & Monitoring	The host institutions provide services and administrative support (space, grants management and human resources) to GA4GH. The Steering Committee members hold teleconferences on a monthly basis. Both the Steering Committee and the management team regularly report to members. Members of the Working Groups initially collaborated through various instruments (from Google docs to Skype calls) but the Alliance has recently developed a common backend that allows members to distribute email, organize online meetings, manage common documents and communicate through web-based conference solutions. Working Groups and demonstration projects report to the Steering Committee.
Decision making Decision making rules Autonomy	GA4GH decisions are taken by the Steering Committee according to majority rule. The Steering Committee decides on the establishment or termination of Working Groups.
Goal setting Strategic decision making	Goals are set by the Steering Committee, upon advice of the Strategic Advisory Board.
Activities	
Project implementation and design	The GA4GH implements and designs three demonstration projects. Demonstration projects share three fundamental characteristics. First, they are low-cost projects. The Alliance has limited funding available and does not fund large-scale initiatives. Second, they have goals achievable in the short term. The scope of the projects is to demonstrate that an open science approach is feasible and have an immediate impact on research activities. Third, they work on neutral, non-contentious data sets (i.e. metadata). In this way, the Alliance is able to gather more actors together, and show the value of sharing a first layer of information. At present, the Alliance has three demonstration projects. Matchmaker Exchange (http://www.matchmakerexchange.org/): this project aims to engage research projects in creating a common platform to facilitate the matching of cases with similar phenotypic and genotypic profiles using a standardized application programming interface (API) and common conventions. Overlapping

	phenotypes of similar genes is fundamental to discover causes of unknown diseases. The project counts between 10 and 15 large datasets among participants. BRCA Project (http://brcaexchange.org/): the project aims to foster collaboration and propose a federated model of data sharing to develop common interpretative tools of data in breast cancer research. The project does not pull data from repositories and does not share them with third parties. It provides an API to recognize, connect and compare data. One of the main difficulties of launching the project was potential competition or interference by GA4GH with established research and practice communities within breast cancer research. The role of GA4GH was not to reshape communities or modify community rules, rather to enable them to accomplish their tasks through technological tools. Beacon Project (https://beacon-network.org//#/): the project aims to facilitate data discovery by scientists through a common query, independently applied by a set of databases. The query does not allow scientists to access data directly, but it allows scientists to interrogate datasets about data. Through the query, scientists can test whether the database contains the data they are looking for. By way of example, the service is designed merely to accept a query of the form "Do you have any genomes with an 'A' at position 100,735 on chromosome 3" (or similar data) and respond with "Yes" or "No".[23]
Sharing Policies	
Data sharing approach	All work produced as part of the GA4GH activities is openly shared, including codes of APIs developed in the demonstration projects. Nevertheless, the Alliance does not require its members to follow rules concerning the sharing of their internal data and information. According to the Alliance, data sharing is based on the respect of data owner's policies. The Alliance does not directly engage in promoting data sharing among members. By not enforcing rules on data sharing, the Alliance aims to attract a large heterogeneity of actors to work together on harmonized approaches and rules. GA4GH supports data sharing by: (1) improving data accessibility through IT tools; (2) developing and suggesting harmonized documents and approaches for data sharing (i.e. consent policy; responsible treatment of data; common vocabulary for data sharing); and (3) supporting collaborative iterations among members. GA4GH tools do not allow researchers to access data. They facilitate data localization and comparison. In GA4GH strategic vision, this approach attracts a larger variety of investors and stakeholders (including private companies) that are willing to develop data sharing enabling tools but not directly engage in common pools of data. *Material sharing*

[23] Source: https://genomicsandhealth.org/work-products-demonstration-projects/beacon-project-0

	The Alliance focuses on accessibility to genetic material. As it is extremely difficult to promote material sharing because of protective national regulation, GA4GH is collaborating with biobanks and biobank consortia to obtain metadata and make those metadata more accessible to researchers. In that way, GA4GH aims at facilitating the localization of genetic material, as a first step to enhance material sharing. GA4GH does not enforce rules concerning material sharing: once the researcher has located the material, it is his responsibility to submit a request to access to it.
Data sharing policies	There are no policies enforced for the sharing of members' internal data and information. All work produced by the Alliance is freely available online.
Evaluation and Findings	
Course of the analysis	It proved difficult to schedule interviews with GA4GH Board and management members. The case study relies on two extensive interviews and documents collected on GA4GH website. All gathered information converge on recognizing accessibility and development of harmonized policies as the focus of the Alliance. It seems that the Alliance has a federative role in the sector, and leverages on this role to promote the adoption of shared practices and conventions in the sector. The Alliance does not pursue any monitoring or enforcement of rules among actors and relies on consensual, collective instruments and processes to pursue its goals.
Key findings	**Accessibility before sharing.** The focus of the Alliance is first to ensure that scientists are able to identify where data are located). This dimension of accessibility is a first step towards data sharing and collaboration. **Showing values.** The demonstration projects are key for the Alliance. They show the values of the Alliance and engage members in common initiatives. The projects strengthen links with research communities in the filed of genomics for human health and gradually propagate the GA4GH approach. Projects have to demonstrate value in the short term to incentivize members, and should be focused on neutral, non-sensitive matters in order to attract a large variety of actors. **Weak rules and large participation.** The Alliance does not set, monitor or enforce rules. It pursues a leadership position in by aggregating all relevant actors. As more and more influential actors adopt GA4GH instruments, others will follow. **Common goals and heterogeneity.** The initial aggregation of a group of actors with a common vision and goals is helpful to gather enthusiasm and energy around the initiative and establish common ground. Once the initiative has developed its goal setting and decision-making procedures, the group can be further expanded to include more heterogeneous actors. **Representation and management.** The structure of the Alliance is a balance between the need to have inclusive, representative governance bodies, and a

	management structure that leads and manages projects and common activities with flexibility (demonstration projects and working groups). A formalized Steering Committee and more informal Working Groups further materialize this balance. **Setting expectations and goals.** The foundational documents designed by the first consolidated group of Alliance members, were instrumental to principled engagement. The documents clearly articulate the scope of the initiative and lead joining partners to level their expectations.
Data Sources	
Interviewees	• GA4GH Director • A Working Group director
Attached references	• GA4GH_White Paper • GA4GH_Working group activities • GA4GH_Mission Rules Principles • GA4GH_Workflow • GA4GH_Beacon Project • GA4GH_Constitution • GA4GH_Governance structure • GA4GH_MatchMaker Exchanger • GA4GH_Road map • GA4GH_Framework for Responsible Sharing of Genomic and Health-Related Data
References	Knoppers, B. (2014). Framework for responsible sharing of genomic and health-related data. *The HUGO Journal*, *8*(1), 3. http://doi.org/10.1186/s11568-014-0003-1 Kosseim, P., Dove, E. S., Baggaley, C., Meslin, E. M., Cate, F. H., Kaye, J., … Knoppers, B. M. (2014). Building a data sharing model for global genomic research. *Genome Biology*, *15*(8), 430. http://doi.org/10.1186/s13059-014-0430-2

Appendix 5. Integrated Breeding Platform

Project Information	
Name	Integrated Breeding Platform – IBP https://www.integratedbreeding.net/
Mission	The Integrated Breeding Platform (IBP) is a downloadable platform conceived to support breeders to "accelerate the creation and delivery of new crop varieties".[24] It provides breeders with an integrated system for the collection, storage and analysis of data in all breeding phases, and by offering technical support, training and community space for communication with experts and other breeders. The platform is offered at different price structures: free for universities, non-profit and governments in developing countries and modular for private companies, government and research institutions in North America, Europe and Australasia.
Field	Food and agriculture
Brief history	IBP was initially developed in 2010 by the Generation Challenge Program (GCP), which was terminated in 2014. 2001 – 2002 Design and development of GCP 2003 - 2004 Launch of GCP and start of the research and funding programs 2004 – 2008 GCP Phase I 2009 – 2014 GCP Phase II 2009 – Transition towards IPB and online launch of the project 2010 – Official launch of Integrated Breeding Platform
Budget and funding source	IBP is currently funded by the Bill and Melinda Gates Foundation, UK Department for International Development (DFID), the European Commission, IFAD, CGIAR, the Swiss Agency for Development and Cooperation. Over the first five years of the platform (2009 – 2014), GCP allocated US$ 22 million to IBP, with financial support primarily from the Bill & Melinda Gates Foundation and from the European Commission and DFID.[25]
Size	The project currently employs 28 as central management staff (excluding employees from partner IT companies and regional hubs employees).[26] The regional hubs acknowledge a similar number of IBP representatives.

[24] Source: https://www.integratedbreeding.net/2/about-us
[25] Source: http://r4d.dfid.gov.uk/PDF/Outputs/GenerationChallenge/8_Integrated_Breeding_Platform_DRAFT.pdf
[26] Source: https://www.integratedbreeding.net/8/about-us/governance-management

	IBP is currently partnering with 4 organizations for software development, and acknowledges over 50 partners that have contributed to the development of the platform.
Location	The central team is located in at CIMMYT Mexico. Regional hubs are currently located in Benin, China, India, Kenya, Nigeria, Senegal, and Thailand. Upcoming regional hubs will be located in Brazil, Colombia, Europe (France), United States, Philippines, South Africa, and Zimbabwe.
User community	Users of the platform are located both in developing and OECD countries. In developing countries, users include CG centers, several national programs and research institutes in countries across Africa and Asia including Ethiopia, Kenya, Senegal, Philippines, Bangladesh, India, China and Thailand. In OECD countries, the platform has been adopted by universities, such as UC Davis and Texas A&M. IBP targets researchers and breeders. Breeders include both new-generation breeders who have been exposed to modern breeding technologies and more traditional breeders seeking new approaches and techniques. IBP has signed service contracts with 2 private commercial companies.
Technology	IBP is providing an integrated breeding platform where users can manage and analyze data. At the moment, the platform is downloadable on computer by users, but future goals include the development of a cloud-based platform to support data storage and facilitate data sharing across users. The platform has been built in collaboration with external IT companies and includes the integration of several tools that are available on the market in open source. Creators are acknowledged on the website.[27] The platform is flexible and can be adapted to different crop-based needs. It allows researchers to choose among a wide set of tools. The platform is developed in open source, leveraging on iPlant and other open source tools that were developed by university and research programs. Nevertheless, access to the platform is subject to a formal agreement with IBP. The platform and the additional tools are free for developing country universities and government research programs. Others can access the platform according to a fee system which is proportionated to the resources available to the organizations. Some of the optional tools of IBP are developed and managed by private companies, and available at the same conditions.
Project Description	
Mission and main activities Goals	The goal of IBP is to provide the technological tools and technical assistance that are needed to enable data sharing, rather than directly create incentives for data sharing or collaboration. Indeed, IBP offers a standalone, integrated breeding management system (BMS) to different types of users, from universities and public research programs to

[27] For the complete list on contributors to the platform and private companies, see: https://www.integratedbreeding.net/232/about-us/our-partners-and-funders

Activities Boundaries	breeders and private companies. The platform is designed to assist with data management in all breeding phases. Breeding includes both traditional and molecular breeding, with a stronger emphasis on the latter. IBP also makes available technical staff to help users with data-related management activities, including data transfer to the platform, data formatting, curation and analysis. IBP also offers training and workshops for current and future users of the platform. While in the first phase the project (2009-2014), most of IBP activities focused on platform development and capacity building, platform dissemination and implementation, as well as user recruitment, are the priorities in the on-going second phase. One element of IBP's mission is to introduce developing country potential users to technology, for activities that are still largely manually performed in developing country realities. Platform development and maintenance, deployment and dissemination of IBP tools and services in the different regions, coordination of partners, including regional hubs, and commercialization of the platform, are IBP core activities. Commercialization and dissemination are widely supported by training activities that allow users not only to implement the platform but to actually be able to leverage on it for their research and breeding purposes. IBP does not provide grants, and does not produce or store data.
History and drivers Foundation of the project Evolution of goals Drivers Lessons learnt	IBP was created within the Generation Challenge Program (GCP). GCP started between 2001 and 2002 and was officially launched between 2003 and 2004. It was designed to leverage on the untapped potential of genetic resources, especially those stored in CG centers, for development research. GCP had a timeframe of 10 years and around US$ 170 million funding (the website reports around $ 15 million per year).[28] More than 200 partner institutions were involved in the program. GCP founders believed that a time-bounded project would have helped to focus on achievement and there was a general acknowledgement that technological changes would have made the project redundant beyond the 10-year life span. The program was initially presented and endorsed by the CGIAR Science Council. GCP aimed to financially support research on specific crops. Funding supported both commissioned research and competitive grants. With commissioned research, GCP requested scientists to develop research plans in specific areas. For competitive grants, scientists proposed their own research projects and GCP assigned funding on a competitive basis. Throughout the entire program, preference was given to supporting grants for developing country scientists. In its last years, around 50% of the budget was devolved to developing country-based projects. In many cases, GCP attempted to promote scientific collaboration between developing and developed countries by pairing scientists with similar research goals. To overcome trust barriers and prioritize capacity building, GCP favored developing country leadership in those projects. In this way developing countries scientists felt their interests were protected and were able to further learn both scientific and leadership skills.

[28] For more information on funding, see: http://www.generationcp.org/network/funders

	Close to the 10 year expiry, the GCP executive team started revising the program to address some of the challenges that had limited the project impact. First, the community-driven approach had resulted in highly fractioned leadership and the project was encountering more and more difficulty in establishing and achieving common goals. Second, without the ability to move forward, the project was losing trust from participants and partners whose high expectations were not fully met. By way of example, although GCP was structured to strongly encourage data sharing among partners, little data sharing occurred. Additionally, much of the data that was shared was of low quality due to lack of standardization and common formatting. In response, GCP switched from a community-driven approach, which was based on institutional rules on representation and participation by partners, to a technology-driven approach. The new philosophy is to support trust and collaboration among actors involved by promoting reliable common technical tools, namely the Integrated Breeding Platform. By developing a common platform, GCP aimed at focusing all efforts towards a clear, common project whose characteristics would have in turn facilitate collaboration and data sharing among other actors in the long run. A critical initial step was the identification of potential users and needs that would guide the design of the platform, including platform functionalities and tools. The GCP committee relied on external consultants, market research and GCP community feedback in order to identify the potential demand among private companies, universities and public research programs. In this heterogeneous context, it was likewise important to set the boundaries of the initiative. Requests from GCP partners, which included actors with different level of capacity and diversified goals, were making increasingly difficult for the newly born initiative to address all needs in an efficient way. It was fundamental for the management to clearly define which activities and functionalities would be excluded from the platform. IBP decided that platform functionalities would only focus on breeding, with complementary education, communication and outreach activities. IBP is not directly involved in research, data analysis or data production. A further challenge was the design of a financially sustainable model. A fee-system was put in place. Access to the platform is generally free, but additional tools and support are priced according to the user financial capability. In practice, most of the BMS is free for developing countries institutions, and revenues are generated by collaborations with developed country research and government institutions and private companies. Engaging with companies remains an actual challenge.
Governance	
Macrostructure	
Membership: Rules & benefits	IBP is not built around membership. Rather, it provides subscribers with services. Subscription does not entail participation in IBP governance. Users can just provide IBP with suggestions or requests concerning the functioning of the platform. At present, most of IBP users are university and government research projects.

Private sector participation Developing country engagement	IBP is making attempts to enlarge its user basis to private actors, especially medium and small companies working in a defined geographical area. Companies with limited capacity to produce data would gain from sharing with a group of other companies. Companies in developing countries with shortage of genetics data and material would also be attracted to IBP, especially to early access to data and pool of data/material. Large companies do not have such a need and already have in-house developed IT systems for data management. At the current moment, two private companies have subscribed to IBP services. IBP has a strong focus on developing countries, as it was in GCP. IBP aims to provide a free infrastructure for advanced breeding, and to empower developing country scientists through collaboration networks around the platform. For this reason, IBP has created the regional hubs and offers free training and collaborative workshops.
Structure Governance bodies Management structure Teams and skills	IBP is designed around two levels, namely a centralized team in charge of coordination, monitoring and strategic decision-making, and a group of regional hubs, which locally support IBP implementation. *Central structure* GCP was led by a representative board, where all stakeholders shared authority and responsibility for decision-making. The board was only partially effective as each member tended to conservatively represent the interests of his/her organization of affiliation instead of prioritizing GCP common goals. The board was directed by a Chair and was responsible for promoting accountability and transparency within the initiative. A Chief Executive was in charge of program management along with an executive board. The executive board was smaller and more effective than the representative board, and members served in their personal capacity. During the transition from GCP to IBP, GCP established a Scientific and Management Advisory Committee to support the IBP team and guide the project. When the initiative was set up, there was a general agreement within the Committee that IBP should be designed with a light governance structure to avoid fragmentation and fast-track implementation. At present, IBP functions with a small Board of Trustees (5-6 members), which is in charge of strategic leadership of the project. Board members include the IBP Director, who leads the Management Team. The Management Team is in charge of all daily management activities, setting goal priorities and decision-making processes. It includes five professional profiles: a commercial manager, a product manager, a capacity development manager, a deployment manager and a technical support manager. There are future plans to provide IBP with additional governance instruments, such as an Advisory Board and a Stakeholder Committee. The Advisory Board would be composed by technical experts from different fields, and would be tasked with scientific advice to the Management Team and the Board of Trustees. Governance documents also envisage the establishment of a Stakeholder Committee, as the project

	grows and the partner portfolio expands. The Stakeholder Committee would guarantee accountability to investors. It would meet on a yearly basis. *Regional hubs* The importance of a regional presence emerged from the GCP experience. Most successful collaborations with developing countries were led by local institutions and deeply embedded in local networks. To replicate this experience, IBP has been designed around regional hubs. Regional hubs are organizations that have formally agreed to collaborate with the project. Hubs are created to facilitate operations at the local level as they represent a first and local-driven interface of the project and are able to tailor the IBP offer to local conditions, including infrastructure and technology absorptive capacity.
Processes	
Coordination Responsibilities Reporting & Monitoring	The Management Team is based in Mexico at CIMMYT. The team is in charge of ensuring coordination of all IBP activities, and directly implements commercialization, IT management, and outreach. The Director of the team is in charge of coordinating management activities with strategic guidance by the Board of Trustees. Hubs receive staff support from IBP and commit themselves to support the adoption of the platform in their region, by providing technical support, capacity development workshops and advice to local breeders. Although hubs are autonomous in their daily activities, they collaborate closely with the central management team, to which they have to report, monthly or quarterly.[29] The central team coordinates the hubs on education and training.
Decision making Decision making rules Autonomy	Decisions – which include new partnerships, adjustments to the platform, commercialization activities, among others - are made on a consensus basis among the six members of the management team. Meetings often include, when necessary, representatives from the IT companies that are collaborating on the project and external consultants. The Director reports to the Board of Trustees. Hubs are autonomous in proposing and organizing activities at the local level. They can report suggestions and problems to the management team, which will address them during the coordination meetings.
Goal setting Strategic decision making	Goals are set by the management team, which receives advice from the Board of Trustees. External consultants, IT companies, and the hubs are consulted during the goal setting process, which also consider community feedback on platform functionalities.
Activities	

[29] It seems / According to one interviewee hubs have been operational since one year; first yearly reports should be received by IBP in these months.

Community building	GCP experience helped IBP to design engagement strategies and actively promote community growth, especially through long term localized training. The first IBP training program was designed as a two-week event to be attended every year for three years (2012, 2013 and 2014). The training covered a wide range of topics, from introductory to more advanced material, such as data management, statistical analysis and molecular breeding. [30] Researchers were allowed to use their own data during the workshop. The program involved researchers, equally distributed between three regions, Western and Central Africa, Eastern and Southern Africa, South and Southeast Asia. 136 researchers completed the program. The long-term group approach to training was key to create a first community around the project and identify 'champions' to support the implementation of IBP within new organizations and represent the platform at the regional level. The program helped researchers to develop a sense of belonging with the community, better understanding the value of the initiative and learning how to leverage on the platform for their own research purposes. It was not possible to replicate the three-year training program due to the high costs. Moreover, while training was an initial strategy to support the diffusion of the platform and community-building, the management team judged commercialization and large diffusion greater priorities for Phase II. IBP focuses on the development of stronger links with universities and other organizations that might be potential users of the platform, partially by leveraging on contacts with researchers who have previously been involved in Phase I or in GCP. Training is still widely provided on site to institutions interested in IBP services. Three deployment teams are in charge of on-site training in three different regions, Western and Central Africa, Eastern and Southern Africa and South and Southeast Asia. Each deployment team is composed of a manager, breeders and data managers. Deployment teams provide on-site training to institutions which are already negotiating the implementation of the platform.
IT management	The development of the platform features among IBP objectives. Platform development includes both daily maintenance and long-term development goals. Feedback is central for assuring a high quality product that meet customers' expectation. In 2015, the feedback system was revised in order to better categorize and prioritize requests from users. With regard to long term goals, IBP is planning to switch from a desktop-based platform to a cloud-based system to facilitate data storage and data sharing.
Commercialization	The long term sustainability of IBP relies on platform subscriptions by private companies and government and research institutions in North America, Europe and Austrasia. A private company with experience in developing country markets is responsible for commercialization. The company receives a baseline fee and commission fees. Commercialization includes comprehensive customer support, including to transfer and format data into the system. Most of those activities are free,

[30] The course was called Integrated Breeding Multiyear Course.

	but IBP and the provider company have recently negotiated an agreement to offer extended support service to user upon payment.
Sharing Policies	
Data sharing approach	Both GCP and IBP have experienced strong reluctance by potential providers to share data. In GCP, data sharing was part of the grant agreement with funded researchers. At the proposal stage, researchers had to present a data sharing plan to which they had to commit when accepting the funding. GCP created multiple incentives to data sharing. First, GCP supported the development of a database, in collaboration with Bioversity International where researchers could store their data.[31] Second, GCP offered a six-month embargo period to allow researchers to publish. The embargo was renewable for additional six months. Third, the grant agreement guaranteed an additional, adequate extension of embargo period in the case of patentable subject matter. Fourth, GCP supported work on ontology development in order to facilitate data standardization. Nevertheless, GCP failed in promoting data sharing among scientists. Most of shared data was of low quality, poorly standardized and not reusable by other researchers. IBP was conceived largely to overcome those issues. IBP is now following a completely different approach with data sharing, by leveraging on individual motivation to share, instead of formal requirements. IBP has not established obligations for data sharing. IBP allows users to freely choose which data they want to share and with whom they want to share them. Data can be kept private or share with their own research group. Nevertheless, future plans foresee to offer premium or discounted services for those researchers who decide to share their data and IBP will provide free support for data management, formatting and curation. This type of support is already available, despite the fact that IBP is not currently offering a facility for data storage. The support team suggests external facilities where data can be stored. While the current version of the platform is a standalone software, the future platform will be cloud-based to facilitate data storage and sharing. The cloud-based platform is mainly targeted for research institutions, but IBP management hopes that commercialization of the platform will result in private companies sharing. In this scenario, IBP does not foresee that all data will be made publicly available. IBP aims to position itself as a "broker of data", by providing the necessary infrastructure and acting as an intermediator among data owners for data sharing.
Data sharing rules	As described in the previous section, IBP does not foresee rules for data sharing. Actors are free to decide if they want to share their data, with whom and under what conditions.
Evaluation and Findings	

[31] Need to find link

Course of the analysis	The interviewees report a similar perspective on IBP and its evolution from GCP. All of them emphasize how IBP has been designed in continuity with GCP's mission of promoting scientific collaboration and research, but has adopted a different approach in order to overcome challenges emerged in the precedent experience – mistrust, fragmentation, lack of shared vision. IBP is centered on the provision of services to build capacity among actors and federate them around a common tool – the Breeding Management System. All interviewees are critical with respect to data sharing. They report how GCP experience has taught that a rule-based data sharing system does not offer adequate incentives to share data. Obligations are not a sufficient condition to stimulate positive behaviors. Hence, IBP will try to leverage on positive incentives to share data, by offering an added value to those actors who are willing to make their data public – i.e. discount, support. IBP hopes that by reducing barriers for sharing through the platform and by offering positive incentives, they will be able to enhance data sharing at the local and global levels. Finally, all interviewees agree on the fundamental role of interaction with users. For this reason, IBP is structured around regional hubs to localize access to services. The offer is thus customized to meet users' needs and to empower them in effective using the platform. Training is crucial to build capacity, which reinforce trust towards the platform and help scientists to collaborate on the same level.
Key findings	**Regional dimension -** The regional dimension is key in IBP. It is a challenge to find a one-fit-all strategy in global organizations that reach out to users with significant infrastructure and capacity imbalances. Hence, working "local" is important to be effective in offering services that effectively empower users. Moreover, local networks are important to create synergies among actors that facilitate the diffusion of the project. **Demand-driven -** Recognizing the demand that emerges from the community is essential both at the design stage of the project and further on to adjust the direction of the project. Developing use cases and user feedback foster the demand-driven approach. **Governance –** Large boards where multiple interests are presented might be little effective in moving projects towards a shared direction. Advisory boards are important to provide strategic advice and stakeholder boards are important to monitor project's results. Nevertheless, small, skill-based teams are more effective in managing the project and help to increase project legitimization within the community. **Incentives for data sharing -** Facilitating data sharing by removing barriers or introducing legal requirement is often not enough to make researchers share their data. IBP is experimenting a way where researchers get actual incentives (i.e. services) to share their data. In the words of one interviewee, this is a more a carrot-driven approach than a stick-driven one, as it was in GCP. The fact that actors are more willing to share if they see an actual value in the exchange is shown also in knowledge-based exchanges. In the experience of IBP, private companies are willing to share

	information about their dataset if that is useful for developing a platform that responds to their needs.
Data Sources	
Interviewees	• IBP Director • IBP Management Team, including private companies collaborating on commercialization and IT management • IBP operative staff • Previous members of GCP Advisory Board
Attached documents	• GCP organizational chart and Board members • IBP Regional Hubs representatives _ Full list • IBP Regional Hubs _ Map • IBP Pricing strategy • IBP Deployment approach • IBP structure • IBP Workflow Slides • IBP Annual Report 2014 • IBP List of tools

Appendix 6. International Rice Informatics Consortium

Project Information	
Name	International Rice Informatics Consortium (IRIC) www.iric.irri.org
Mission	IRIC is an international consortium that produces and facilitates accessibility to rice genomics data and enhances information exchange among the rice research community.
Field	Food and agriculture
Brief history	IRIC was formally launched in January 2013. The initiative was proposed and led by the International Rice Research Institute (IRRI) as a part of the Global Rice Science Partnership (GRiSP). At the current stage, IRIC is still under development. The management team and the advisory committee are discussing future steps by the initiative, including the definition of shared resources and membership rules.
Budget and funding source	IRIC is financially supported by IRRI and GRiSP, and by membership funds. The budget is discussed and approved on a yearly basis. IRIC interviewees consistently report difficulties in adequately financing project activities (i.e. gene sequencing and phenotyping data collection) and the need for further resources to expand the management team. The Advisory Committee is considering possible funding options, such as leveraging on private sector membership or collaborating with members on research projects.
Size	IRIC management team includes 4 members who are in charge of IRIC scientific and organizational tasks. The development team counts 10 to 15 members.
Location	IRIC team is located at IRRI's headquarters, in The Philippines.
User community	IRIC initial members are: the Arizona Genomics Institute, Cornell University, the International Center for Tropical Agriculture (CIAT), the International Rice Research Institute (IRRI), the Japanese National Institute of Aerobiological Sciences and the Genome Analysis Center (UK). At presents, IRIC counts around 10 members, including two from private sector. The community utilizing IRIC data is much broader and includes institutions from both developing and OECD countries, as well as public, non-profit and private sector organizations.
Technology	IRIC provides a web-based interface to access rice data along with basic analysis tools. The technology part of the project is still under development but is not central in the strategy. IRIC aims to develop an international platform for accessing rice data worldwide, rather than a technologically advanced system for

	data analysis and management. Software tools available under IRIC include mainly APIs.
Case Description	
Mission and main activities Goals Activities Boundaries	IRIC's mission is to promote rice biodiversity by increasing the ability of scientists and breeders to access extensive rice genomics data, for scientific research and innovation. IRIC is organized around three main sets of activities. First, IRIC aims to sequence rice material that is currently conserved in genebanks. Genomics data that are produced through IRIC activities are made freely available on IRIC's web-based platform. At present, IRIC has made available datasets from the 3,000 Rice Genome Project developed by IRRI. The project foresees the production of associated phenotypic data. Second, IRIC plans to integrate and manage available public rice genomics datasets to facilitate accessibility. By integrating and making accessible a large variety of datasets, IRIC aims at promoting technical standards and operability. Third, IRIC promotes a community of researchers that utilizes and contributes to the high quality data stored in IRIC's platform and collaborates on projects that might lead to the production of new outcomes to be shared with the community. IRIC does not promote a research agenda and does not support the development of a comprehensive data analysis and management platform.
History and drivers Foundation of the project Evolution of goals Drivers Lessons learnt	IRIC is a recent initiative; its structure, goals and resources are not clearly defined and IRIC faces the challenge of positioning itself within the rice community. The project started in 2013 and the first steps were the definition of the consortium agreement, the development of technical aspects of the web portal and standards for metadata and interoperability across datasets. After this initial technical phase, IRIC has invested most of its resources (financial, human) in building a common resource, the 3,000 Rice Genome Project data. The curation of those data required almost two years to make data fully disclosed and accessible to the public. IRIC's targets for the following years are the integration of (1) phenotyping data in the 3,000 Genome Project; and (2) other publicly available rice datasets into the platform. Both targets require investment to collect and curate data. Data are to be re-mapped into a standard format consistent with IRIC's platform. The collection of phenotyping data requires funds for field trials and development of common standards across projects to collect data. The attention on phenotyping data is high given the added value that phenotyping data could bring to IRIC datasets. Phenotyping data are more complex to be produced and collected because require financial resources, time, genetic material (i.e. the seeds sample) and breeding labor, but they are essential to utilize and

	fully understand genotyping data. Private sector actors are willing to pay to generate and access phenotypic data. IRIC plans to leverage on IRRI's access to rice genetic material and networks around the world (i.e. in Africa, Southeast Asia and China) to collect data. IRIC aims to coordinate efforts towards common sets of accessions in multiple locations over multiple years. The implementation of those activities – integration of public datasets and phenotypic data – is crucial for the long-term success of the project. IRIC management recognizes that it has not yet developed a common resource able to attract the rice community towards its platform. Membership rules do not yet define clear benefits for IRIC members as compared to non-members. By collecting further data and develop further analysis on them, IRIC hopes to build a better resource that will attract a greater community around the project and to be able As it aspires to offer its own genotyping and phenotypic data, IRIC's management is concerned over other, that better financed initiatives might take over IRIC data – which are free and publicly available - and develop a similar infrastructure. The key question is to figure out the good mix of funding that IRIC might collect to achieve its goals. In this case, funding might come from private sector and grants for phenotyping activities.
Governance	
Macrostructure	
Membership: Rules & benefits Private actor participation Developing country engagement	Membership to IRIC is open to any public sector organization, non-governmental organization or private company. Membership is divided into two categories: (1) private sector and (2) public sector. Private sector membership is subject to a US$20,000 fee, while public sector members are free to determine their contribution. All members have to sign a formal agreement to join IRIC. Reportedly, there are 3 to 4 private sectors members and 8 public sector members at present. A central open question in the development of IRIC is the design of membership advantages. Although a large community of researchers is using IRIC data, formal membership is limited. At presents, there is no incentive for organizations to join IRIC because as non-members they can access to the same resources. IRIC is considering additional data tools or advanced analysis data for members only. Additional advantages might include early access to data.
Structure Governance bodies Management structure	IRIC is governed by an Advisory Committee and a management team. The Advisory Committee is composed by 6 members, which include 2 private sector members, 3 public sector members and an IRRI representative. The Advisory Committee is in charge for three years and is elected by IRIC members. The IRIC coordinator is the secretary of the Advisory Committee. The Committee

Teams and skills	is responsible for reviewing membership applications and launching new partnership, reviewing the budget, monitoring and evaluation. IRIC is managed by a project coordinator who is appointed by the Director General Director of IRRI as full-time staff. The management team is small (3 to 4 members) and is supported by a larger IT development team (10 to 15 members). The need for a project administrator and technical staff to support the management of the project and its scientific development, is recognized.
Processes	
Coordination Responsibilities Reporting & Monitoring	Coordination happens through weekly management meetings at IRRI's headquarters. The Advisory Committee meets yearly in person and online 3 to 4 times per year. The Advisory Committee oversees all management activities and coordinates initiatives with external partners.
Decision making Decision making rules Autonomy	IRIC is an autonomous entity, although IRRI is involved in strategic decisions. Strategic decision making is coordinated by the Advisory Committee members.
Goal setting Strategic decision making	Strategic and scientific directions of the project are established by IRRI (as lead center of GRiSP) in consultation with IRIC members. All IRIC members can provide suggestions on the future direction of the project.
Activities	
Outreach	IRIC is engaged in presenting its platform, data and activities in conferences and events in order to attract and connect with potential stakeholders. IRIC also engages in collaboration with external partners. Some partners, such as CIAT, CAOS and NIAS Japan, have become IRIC members over time. Collaboration starts through discussions with various teams and the combination and integration of complementary interests between IRIC and the partner. Collaboration generally includes both a technical and a research component. For instance, collaboration with CIAT is for the development of an API for accessing the 3,000 Rice Genome Project data and integrating the data with CIAT's datasets. The research interest of CIAT is the structural variations that can be traced into the dataset. Collaboration with CAOS includes support on population genetics analysis whose results will go into IRIC platform. NIAS, instead, collaborate with IRIC for curating gene names to be used both on NIAS projects and IRIC's platform. Collaboration is made possible as there is a complementarity between the needs of the two partners. The 3,000 Rice Genome data allows IRIC to attract partners that

	are interested in developing new tools for accessing and leveraging those data. Collaboration are coordinated by IRIC and IRIC team actively interact with all partners.
Sharing Policies	
Sharing approach	At the current stage, all IRIC data are freely and publicly available. This approach is still possible since all available data are directly produced by IRIC in collaboration with IRRI and are by default publicly available. Indeed, all data stored in IRIC are publicly accessible in other public datasets, such as Amazon Public Data, and not just through the platform. The added value by IRIC is in the tools that it offers for further analysis and integration. IRIC aims to create a community of rice researchers and breeders that utilize IRIC data and contribute with their own data to the platform. The proposition is to create a community that utilizes the public resources currently available on the platform and to engage with different actors to find complementarities that might lead to new research projects. IRIC has not yet discussed with members requirements and rules for data sharing, and generally apply the Toronto Agreement to members who decide to share their data (see below). Members who want to share data include scientists that have received grants or publish in journals that require data to be open. IRIC encourages reference by rice scientists in publications.
Sharing rules	IRIC applies the Toronto Agreement. The Toronto Agreement recognize data producers the right of first publication and data user the right of accessing public utility dataset in the shorter delay possible. IRIC also encourages members to share datasets for private use only. This policy allows scientists: (1) to examine data and develop analyses for internal use; (2) to start with analysis while complying with embargo periods; (3) to look at dataset structures and harmonize them. IRIC currently holds information on data contributors and users. The tracking system will be maintained in the next phases for IRIC to monitor the functioning of the platform.
Evaluation and Findings	
Course of the analysis	Interviewees are generally aligned, with three recurrent themes that emerge from all interviews. First, all interviewees insist on the need to design clear resources that are offered by the project. Proposals include the integration of public datasets and phenotypic data. Second, they recognize that membership benefits should be defined. IRIC has a limited number of members but a large community that utilizes its data. A key question is how to engage with the larger community and build a more stable relationship with users.

	Third, interviewees highlight the need for further resources to enlarge the management teams and undertake data curation and collection.
Key findings	**Defining the resource is key.** IRIC is facing difficulties in attracting a stable community around the project. Interviewees recognize that developing a better common resource is key to engage with the community and consolidate relationship. In the absence of a clear definition of the resources and their benefits, users cannot engage in the project because they cannot understand its value. **Phenotypic data.** Phenotyping data would be a unique resource for members of the project. This uniqueness will create advantages in applying for funding for data collection activities and will attract additional members from both private and public sector. Rice phenotyping data might represent the niche in which IRIC could position itself in the rice community. **Time.** The slow growth of IRIC is in aligned with findings from other projects. In most of the cases, the first 2 – 3 years are critical to define the project resources and activities. None of the project has been built in a shorter timeframe. **Management capacity.** The governance structure of IRIC is still simple as the project has still few members. Nevertheless, IRIC recognizes that the project expansion requires first of all a larger management team that is actually able to follow and implement all activities. Management capacity is key to let the project grow and manage the community.
Data Sources	
Interviewees	• IRIC management team members • IRIC advisory board members
Attached documents	• IRIC_Guidelines for partnership in IRIC
References	Li, J.-Y., Wang, J., & Zeigler, R. S. (2014). The 3,000 rice genomes project: new opportunities and challenges for future rice research. *GigaScience*, *3*, 8. http://doi.org/10.1186/2047-217X-3-8 McCouch, S. R., McNally, K. L., Wang, W., & Hamilton, R. S. (2012). Genomics of gene banks: A case study in rice. *American Journal of Botany*, *99*(2), 407–423. http://doi.org/10.3732/ajb.1100385

Appendix 7. Methodology

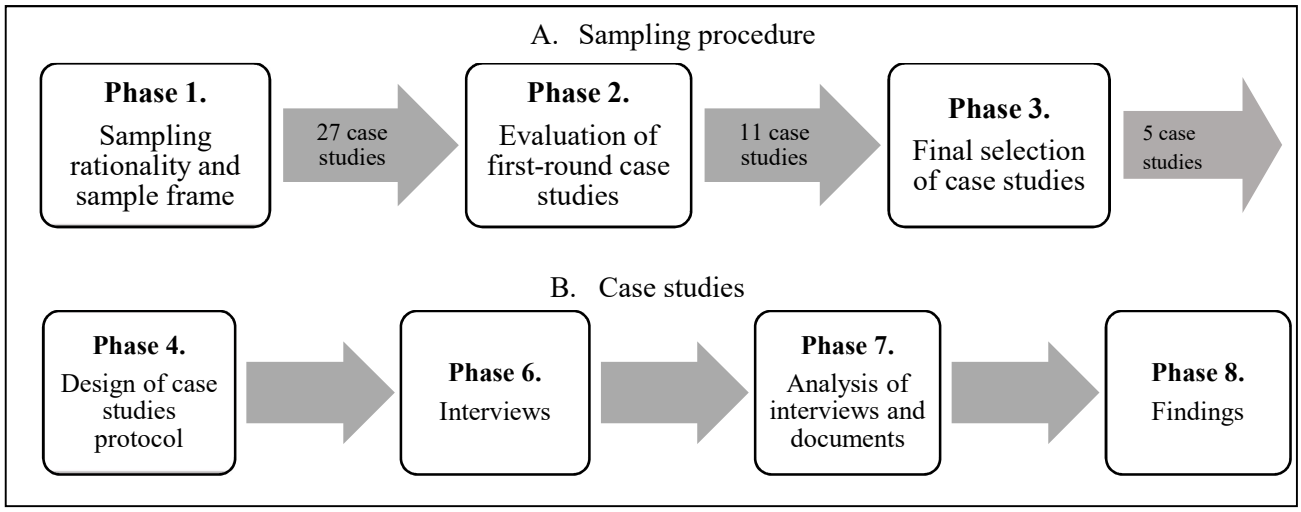
Overview

This part of the document is to illustrate the case study methodology applied in the research project. At the preliminary stage, the research team extensively worked on the development of criteria for case study selection and reviewed relevant literature for the project. While literature review provides a starting point for the investigation, the lack of in-depth, theoretical research in this area requires a more detailed assessment of governance, drivers and data sharing approaches in different contexts. Case study methodology is suitable when the research question is broadly conceived, complex contextual factors are likely to play a significant role in explaining outcomes, the research aims to explore an undertheorized contemporary phenomenon, and when appropriate analysis requires multiple sources of evidence (Eisenhardt & Graebner, 2007; Yin, 2011). Moreover, case study methodology allows a deep understanding of the dynamics that take place within a single setting (Eisenhardt, 1989) and eventually compare findings across multiple settings (Yin, 2011). In particular, case study methodology fits well in studies where the research question aims to "illuminate a decision or a set of decisions: why they were taken, how they were implemented, and with what results" (Schramm, 1971, p. 4).

For this research project, we adopted a holistic, multi-case approach in order to enable exploration of differences within and between cases, and replication of findings across cases. We compared and contrasted collaborative research programs that have different governance structure, participants, rules in use and goals. Careful case selection is essential under this research scenario, so to predict similar or contrasting results across cases (Yin, 2011). We identified programs using a theoretically developed sampling rationale in which conceptually justifiable reasons are the basis for the selection (Glaser & Strauss, 2012).

While we detail the methodology in the next paragraphs, figure 1 provides an introductory overview of the different phases. In the first phase, we built our sampling. In line with the objectives of the study, our criteria emphasize factors that might significantly affect the governance of large-scale genomics projects and d initiatives. Based on selected criteria, we evaluated nineteen cases in food and agriculture research and seven cases in the human health sector. In the third phase, we invited Project Managers and Executive Directors of twelve selected initiatives for an exploratory interview. The interview focused on governance structure, management, data sharing mechanisms and incentives, and critical factors for success of each organization. At the end of the interviews, we selected six organizations for our research, three in food and agriculture research, one in human health and one in science. Part B of the methodology describes data collection and analysis.

Figure 1. Methodology phases: Overview



Sampling procedure

When case study methodology is used to understand and compare social phenomena, setting the criteria for the selection of case studies is a critical step of the research process (Stake, 1995; Yin, 2011). Applying a replication logic might not be adequate, as researchers might prefer maximizing the unique learning contribution of each case and build the most effective structural representation of the phenomenon of interest (Stake, 1995). According to Patton (1990), "the logic and power of purposeful sampling lies in selecting information-rich cases for study in depth. Information-rich cases are those from which one can learn a great deal about issues of central importance to the purpose of the research, thus the term purposeful sampling" (p. 169). For this study, the sampling has been developed in three phases: (1) sampling rationality and sample frame, (2) evaluation of first-round case studies and (3) final selection of case studies. Each phase has been informative for the following one, and the research team have progressively updated the selection criteria according to preliminary findings.

Phase 1. Sampling rationality and sample frame. The sample frame has been designed by searching for relevant initiatives on the Web and on academic articles, and by consulting with professionals and academics in the genomics field. We did not aim at building an exhaustive list, but wanted to identify most relevant initiatives according to the selected criteria. The initial list counted nineteen cases in food and agriculture research, which we further integrated with seven cases in human health and science research, as we moved forward in our research. The complete list of cases is presented in table 1.

Table 1. Summary table of case study selection

#	Project name	Contacted for interview	Interviewed	Selected for in-depth case study
1	Seeds of discovery	**Yes**	**Yes**	No
2	Centre for Integrative Legume Research	No	No	No
4	Soybean Knowledge Base	No	No	No
5	International Rice Research Institute	**Yes**	**Yes**	**Yes (only IRIC)**
6	Genome Canada	**Yes**	No – no reply	No
7	Genesys	No	No	No
8	Cacao Genome Database	**Yes**	**Yes**	No
9	Cassavabase / Nextgen Cassava	**Yes**	**Yes**	No
10	Global Rice Science Partnership	No	No	No
11	Gramene	No	No	No
12	Hap Map	No	No	No
13	Integrated Breeding Platform	**Yes**	**Yes**	**Yes**
14	Structural genomics consortium	**Yes**	**Yes**	**Yes**
15	Tree Genes	No	No	No
16	Germinate	No	No	No
17	International Barley genome consortium	No	No	No
18	Eurisco	No	No	No
19	Iplant	**Yes**	**Yes**	**Yes**
20	GOBII	No	No	No
21	Open Science Grid	**Yes**	**Yes**	**Yes**
22	Global Alliance for Genomics and Health	**Yes**	**Yes**	**Yes**
23	SciCrunch	**Yes**	**Yes**	No
24	Project RedCap	**Yes**	No – Refused	No
25	VIVO	No	No	No
26	European Bioinformatics Institute	No	No	No
27	Canadian Open Genetics Repository	No	No	No

Phase 2. Evaluation of first-round case studies. To evaluate the case studies, we relied on five criteria that we theoretically developed according to the results of the literature review. In particular, we focused onfactors that might influence the governance and management of a large-scale genomics project or initiative in the PGR field. Criteria and metrics used are explained in the table 2.

In this second phase, we selected cases that captured the greatest variation in the key areas of the study (Patton, 1990) and cut out initiatives too similar to each other. Detailed information on each initiative was collected through project or initiative websites and documents available

online (e.g. internal documents, gray literature and academic papers). Several initiatives in the initial list met the criteria and varied in ways important to satisfying this study's objectives. In general, we excluded initiatives that were predominantly academic, while favoring initiatives that involve actors from different sectors (especially from the private one) and demonstrate innovative approaches to data sharing. As shown in the table 1, we invited twelve initiatives to the first round of interviews.

Table 2. First evaluation of case studies: selected criteria

1. Resources Characteristics

Criteria: Projects and initiatives in which scientists and other actors exchange and use genetics data and material.
Rationality: While both data and material sharing might be restricted by legal constraints (Chokshi et al., 2006; Contreras, 2014), material sharing is more exposed to legal burdens at the national and international level (Welch & Louafi, n.d.).

Metrics:

Flows	
Data	Does the project provide access to data?
Material	Does the project provide access to material?

2. Accessibility, Terms of Contribution and Terms of Use

Criteria: Projects and initiatives in which community members freely access and use data and material, and where they are encouraged to contribute to the common pool with their data and material. Projects and initiatives with a clear data access, use and contribution policy have been preferred.
Rationality:

- **Access policies:** Different legal contexts might affect and limit the opportunities for data and material exchange. For instance, IP rights might limit the use and diffusion of data and might prevent actors from freely sharing their resources with community members (Chokshi et al., 2006, 2006; Contreras, 2014; Eckersley et al., 2003; Kosseim et al., 2014; Welch & Louafi, n.d.). As legal issues might affect the data and material access and use, projects or initiatives that aim to promote data sharing design their own policy to facilitate data and material exchange among community (Elta Smith, n.d.; Knoppers, 2009; Kosseim et al., 2014).
- **Data production:** When data and material are collectively produced, it is easier to establish data sharing rules than when data and material are individually collected and produced. In this latter case, researchers own rights on data and material and might seek for more recognition and reward from data and material use (Chokshi et al., 2006)

Metrics:

Access policies	
Open	Access to data and material is open and free. Users might be required to register or agree with the use/contribution policy.
Specified access	Access to data and material is reserved to members.
Data policies	
Users	There are rules for data and material use.
Contributors	There are rules for data and material contribution.
Data production	
Internal	Data are produced by the partners or members of the project or initiative.

External contribution	Data are pooled by members or by members and external users.
Aggregator of public data	Data are aggregated from external datasets, i.e. public datasets.

3. Goals

Criteria: Projects and initiatives that aim to integrate diversified goals (i.e. research, innovation and data sharing) or whose goals have evolved over time to integrate multiple interests.
Rationality: Goals diversity might negatively affect individuals and organizations motivation for sharing and collaborating, while overlapping goals might promote interaction among members and support individual motivation to participate in the initiative (Foster & Sharp, 2007; Powell, White, Koput, & Owen-Smith, 2005; Strandburg, Frischmann, & Cui, 2014).

Metrics:

Goal typology	
Database	The project or initiative promotes access to data and material produced by members, other institutions or uploaded by contributors. The project does not promote its own research agenda.
Research project	The project or initiative pursues its own agenda or financially supports research projects.
Platform	The project or initiative provides an open platform for data and material storage, management and sharing.

4. Discipline and sector diversity

Criteria: Projects and initiatives that include stakeholders from diverse sectors, including public, non-profit and private sector.
Rationality: Actors with the same institutional background or from the same scientific field are more likely to share codes and paradigms that facilitate data and material exchange and collaboration (Dove, Faraj, Kolker, & Özdemir, 2012; Möllering, 2005; Tasselli, Kilduff, & Menges, 2015; Tsai & Ghoshal, 1998). Diversity across disciplines and sector might stress differences across community members and induce conflicts and tensions. Diversity might also reduce trust within the community, thus affecting data and material exchange (Knoppers, 2009; Kosseim et al., 2014; Tsai & Ghoshal, 1998).

Metrics:

Sector diversity	
Public	Public institutions are part of the project or initiative.
Private	Private institutions are part of the project or initiative.
Non profit	Non-profit institutions are part of the project or initiative.
Research	Universities and research centers are part of the project or initiative.
Geographical diversity	
Scope of the project	Geographic area to which the project or initiative aims at providing benefits
Stakeholders nationality	Main stakeholders nationality
Developing countries & breeders	
Developing countries	The project or initiative collaborates or involves actors from developing countries
Breeders	The project or initiative collaborates or involves breeders in training and education activities

5. **Information Technology tools**

Criteria: Projects and initiatives that offer innovative IT tools for data analysis, storage and management will be preferred, assuming equal evaluation in other criteria.
Rationality: Genomics research is increasingly data-intensive. Scientists need tools for data analysis, storage and management (Bouffard et al., 2010; Schroeder, Gonzalez-Perez, & Lopez-Bigas, 2013). Accessibility to technical tools might be an incentive for data sharing and remove significant barriers.

Metrics:

Platform characteristics	
Access to germplasm data	The project or initiative provides access to public datasets
Data management tools	The project or initiative provides tools to manage data
Web based / downloadable platform	The platform can be access only online / The platform has to be downloaded and installed on the user's computer.
Data upload	Users can upload and manage their own data.
Data policy	Users are allowed to choose with whom and to what extent share their data.
Storage facilities	The platform provides storage facilities for users.
Visualization tools	The platform provides visualization tools.
Open source	The code of the platform is open source.

Phase 3. Final selection of case studies. In phase 3, we invited the Executive Director or Project Manager of twelve selected project or initiative for a phone or Skype interview. Each interview lasted around 60 minutes. Ten projects and initiatives accepted our invitation, while one refused and one did not reply to the invitation. The interviews aimed to gain a more nuanced understanding of the cases and identify potential interest for the study (Herbert J. & Irene S., 2004). We developed a semi-structured interview protocol for the interview. The protocol is shown in Table 3. Semi-structured interviews are an appropriate method to explore research topics as they allow to better understand the context, deal with the complexity of multiple and conflicting themes and pay attention to the meanings and interpretations of each actors (Herbert J. & Irene S., 2004). Our questions aimed to understand the governance structure, management processes, data exchange and use mechanisms, and critical factors for the success of the initiative.

Table 3. Interview protocol

Q0: Context and overview

1. What was the prevailing situation before the initiative? What are the reasons that have led to the establishment of the initiative?
2. How diverse are the partner involved in the initiative?
3. How have main common problems been identified?

Q1: Access, terms of contribution, terms of use of data and material

1. How open or close the system is? Who has access to data? What are the conditions to access?
2. What community structures are in place? How are actors connected?

3. What incentives (data, information, knowledge, materials, and money) are in place to motivate actors to contribute to the system?
4. What is the organizational capacity to obtain alternative inputs in case of defection of members?

Q2: Governance

1. Who is in charge of the system?
2. How diversity is reflected in governance structure (in terms of status, disciplines, actors…)? What are the criteria for the exclusion/inclusion from the decision-making processes?
3. What are the formal rules that lead the project? Are there norms that guide access, use, distribution, exchange, and contribution of data and material? If yes, can you briefly explain them?
4. How would you define the level of trust/social capital among the partners / collaborators of your project?
5. How would you define the level of transparency and exchange of information among members? How is communication between actors? How are individual connected to each other?
6. What process of selection of newcomers? To what extent people can voluntarily enter or exit from the network?
7. What type of monitoring systems are in place? How would you define the degree of centralized supervision and coordination in your organization? And the level of discretion (self-management rules/autonomy)?
8. Are there sanctions for individuals abusing the use of data and material, and if so, how are sanctions imposed?
9. Are there differences in terms of sanctions and monitoring systems, by major sector – public, private, non-profit?
10. How do you manage to coordinate the key stakeholders?

Q3: Management

1. How is the program supported financially? Where is it located? How is it organized?
2. What resources – human, financial, institutional – are required to run it?
3. How do you manage (1) socialization processes? (2) recruitment activities? (3) communication activities?

Q4. Effectiveness

1. How effective is the system in terms of frequency and quantity of exchange of data and material?
2. To which extent participation and collective action advance or obstacle desirable outputs?
3. What returns are provided from participating to the organization? Who are the main beneficiaries? To which extent are returns provided?
4. How effective are those benefits in increasing the level of participation?
5. How would you define the willingness to sustain the program by the members?
6. To which extent does the program produce key products, such as: academic papers, new sources or quantities of data, innovations, applications, and so on?

Q5: Constraints, gaps and challenges

1. What gaps would identify in the structure, the processes or outcomes?
2. What are your challenges for the future? What needs should be addressed in order to bring the current activities and short-term outcomes more in line with long-term goals and aspirations?
3. What type of strategic planning or expert assistance would you seek? How would guidance by experts help you in this transition?

For the final selection, we followed a maximum variation sampling strategy, according to which we "purposefully pick a wide range of variation on dimensions of interest" (Patton, 1990). Selected organizations, thus, include:

1. Field. At least one case for each field - Food and Agriculture, Human Health, and Science - with a preference for cases in food and agriculture research.
2. Private actor engagement. We picked cases with low – medium – high involvement of private actors. Involvement was ranked based on:
    a. Number of private actors;
    b. Relevance of their role (i.e. are private actors involved in decision-making?).
3. Heterogeneity of actors involved. Projects and initiatives that engage with diverse actors and organizations. We evaluated:
    a. Sector diversity;
    b. Capacity diversity;
    c. Extent to which diversity is represented in governance bodies.
4. Involvement of members in decision-making. Projects and initiatives that engage their members into governance bodies.
5. IT infrastructure complexity. Projects and initiatives that invest in technology innovation and combine scientific research with information technology development.
6. Data sharing policies. initiatives with different types of data sharing policies, ranging from centralized policies imposed at all members of the organizations to decentralized policies, where members are free to decide to whom they share their data.
7. Developing country involvement. At least two initiatives that engage with researchers, organizations, or governmental bodies from developing countries.

Table 4 summarizes the final selection. iPlant and the Integrated Breeding Platform (IBP) were selected because of their focus on information technology and community development through shared infrastructure. IBP is also an important case for understanding developing country involvement and evolution of goals, because of its earlier incarnation as the Generation Challenge Program (GCP). The Structural Genomics Consortium (SGC) is a private sector-driven initiative which allows study of company interests and incentives for participation in collaborative initiatives. Open Science Grid (OSG) was selected because of its innovative approach for sharing of computational resources The Global Alliance for Genetic Health (GA4GH) is a relatively large, heterogeneous initiative that aims to design harmonized approaches and policies to promote voluntary, secure and responsible sharing of health and clinical genomics data. Finally, IRIC is the most recently established initiative aiming to develop a data and software resource for the rice research community. It provides means of understanding the early development of an initiative and provides a reference for comparison with other cases.

Table 4. Final selection criteria

Project name	Private actor engagement	Heterogeneity of actors involved	Involvement of members in decision making	IT infrastructure complexity	Data sharing policies	Developing country involvement
Integrated Breeding Platform	Medium	Medium	Low	High	Decentralized	Yes
iPlant	Low	Medium/High	Low	High	Decentralized	No
Open Science Grid	None	Low	Medium	High	Decentralized	No
Structural Genomics Consortium	Very high	Medium	High	Low	Centralized	No
Global Alliance for Genetics and Health	High	High	High	Low	Centralized and decentralized	Yes (TBC)
IRIC	Low	NA	Medium	Medium	NA	Yes
Seeds of discovery	Low	Low/Medium	Low	Medium	Centralized	Yes
Cacao Genome Database	High	Low	High	Medium	Centralized	No (Yes)
Cassavabase NextGen Cassava	Low	Low	Low	Medium / High	Centralized	Yes
SciCrunch	Low	Low	Low	Medium	Decentralized	No

Note: the table represents the evaluation of the cases **before** completing all the interviews. Thus, errors in the evaluation of the criteria might be due to the limited data collection.

Case studies

For our study, we apply a holistic, multiple-case design (Yin, 2011). Multiple-case design is useful when researchers aim at exploring different settings and they have expectation of contradictory results depending of the same variables. Our unit of analysis is the single organization. Our case studies have descriptive and explanatory nature.

Phase 4. Case study protocol. Qualitative case study methodology facilitates exploration of a phenomenon within its context using a variety of data sources. This ensures that the issue is not explored through one lens, but rather a variety of lenses which allows for multiple facets of the phenomenon to be revealed and understood (Baxter & Jack, 2008). Based on such methodology, the research team decided to rely on multiple data sources, such as documentation and interviews. The case study protocol (see Appendix A) is useful to structure and summarize the data collected and provide a general understanding of the case under investigation (Yin, 2011).

Phase 5. Interviews. Interviews in the context of a case study research are more like to guided conversations than structured interviews with a rigid sequence of questions and answers (Yin, 2011). For the case study interviews, we follow the same protocol of the first round and we integrate questions after each interview in order to make sure to cover all aspects of the project we are interested in. We interviewed between 4 to 5 people for each case. In most of the cases, we interviewed relevant internal actors, identified through the organizational chart on the initiative's website and by asking suggestions to our first interviewees. In few cases, suggestions have been provided by other actors in the field who were familiar with the project. In some cases, we interviewed external actors. External actors have mainly been indicated by other interviewees. In this way we reduced the risk of selecting irrelevant users or minor partnerships. In few cases we selected external actors by looking at the project or initiative website.

Interviewees have been chosen in order to include individuals who have a relevant role in the governance or management of the project or initiative. Interviews generally include:

- Leadership & Advisory board: at least one representative for each of the main governing bodies (i.e. executive team, advisory board, working groups and any others). We privileged actors covering higher positions within those bodies (Chair, Executive Director, President, and Coordinator, among others).
- Strategic decision-making: between two-three persons in the top management team of the initiative. We aimed at including managers that are in charge of (1) scientific and technical support; (2) community development, memberships and communication, especially with developing countries; (3) commercialization or relationships with private partners.
- Operations and management: at least one person among the operational management team of the initiative. This might include interviews with the project coordinator(s) or officers.
- Users and / or members of the initiative: in some cases, we were able to talk with users or members of the initiative. Users or members were able to provide additional information about the benefits that they derive from the process and their involvement into the community and decision-making processes.