

Introduction: Towards Responsible Plant Data Linkage



Sabina Leonelli and Hugh F. Williamson

Abstract This chapter provides a framing for this volume by reviewing the significance and the organisational, technical and social opportunities and challenges related to plant data linkage. We review what “responsible practice” means in relation to the plant environments being documented, the infrastructures used to circulate data, the institutions involved in data governance and the communities involved in plant data work. We show how, across these domains, responsible plant data linkage involves consideration of technical, legal, ethical and conceptual dimensions, thereby: (1) creating and maintaining digital infrastructures, technical standards and discussion venues focused on critical data reuse; (2) developing adequate legal and institutional frameworks that work transnationally; (3) identifying and implementing guidelines for what constitutes acceptable data use, together with systems to monitor and allocate responsibility for breaches and mistakes; and (4) considering the variety of views on what constitutes agricultural development in the first place and how plant research can sustainably, reliably and responsibly contribute to achieving food security. The production of sustainable, responsible and reliable agricultural solutions in the face of climatic and political change depends on the flourishing of transnational, interdisciplinary collaborations such as those represented in this volume.

1 Introduction: Why Care About Plant Data Linkage

Global challenges such as climate change and the needs of a rapidly growing population have led to the emergence of new priorities in plant science and agricultural research. There is increasing interest in crops from the Global South that have been relatively neglected in previous agricultural development schemes, especially

S. Leonelli (✉) · H. F. Williamson

Exeter Centre for the Study of the Life Sciences (Egenis), University of Exeter, Exeter, UK
e-mail: s.leonelli@exeter.ac.uk; h.williamson@exeter.ac.uk

© The Author(s) 2023

H. F. Williamson, S. Leonelli (eds.), *Towards Responsible Plant Data Linkage: Data Challenges for Agricultural Research and Development*,
https://doi.org/10.1007/978-3-031-13276-6_1

those perceived to have less commercial value yet remain of great importance to smallholders. Improving research and understanding on heritage and orphan crops, as well as the wider set of crop varieties, are now recognised as important goals (Ribaut & Ragot, 2019). Given rapidly changing environmental and climatic conditions, deepening our understanding of genotype by environment interactions (GxE) also constitutes a key goal, especially the impact of environmental stressors on phenotypic traits.

The acquisition, curation and interpretation of data about plants, their environments and their human consumers play a central role in these efforts. Research in the plant sciences is marked by a high volume and heterogeneous range of data formats and sources, including quantitative, observational as well as imaging data generated by field trials, breeders, agricultural machinery, agribusinesses and seed distribution companies, publicly funded scientists, and national/regional institutions. These data are certainly “big”, and yet they are neither easy to access nor easy to use. Making these data *accessible* to those who may wish to analyse them is proving an intricate challenge, with large efforts around the world devoted to expanding data access for research purposes and complications emerging from the privatisation and commercialisation of such data. An even greater challenge is to foster fruitful data analysis and interpretation given the countless forms of expertise, goals and perspectives involved: in other words, to make those data *usable* despite their heterogeneous provenance and even more heterogeneous re-purposing, and ensure the reliability and effectiveness of the resulting knowledge, technologies and interventions.

This is why *data linkage*, understood as the ability to connect and jointly analyse diverse datasets, has emerged as a key global challenge for agricultural research and development in the twenty-first century. Agriculture has long depended on the exchange of biological materials and knowledge, but the opportunities for data collection, dissemination and analysis opened up by computational technologies have dramatically expanded the potential of data-intensive research in this domain. Linking heterogeneous data helps to conduct analyses that address the multiple scales that impact plant growth and traits, from the molecular through the physiological to the social and environmental. This in turn facilitates understanding of the complex, scale-spanning phenomena underpinning sustainable food production and environmental management under rapidly shifting climatic and socio-political conditions.

The roots of the multiple challenges involved in linking data of relevance to agriculture are cultural. The landscape of plant data production, circulation and use is marked by the encounter between different *cultures of data exchange*, which in turn creates substantive technical, legal and social challenges to data linkage. At the scientific end of this spectrum, plant science has long sat at the intersection of the laboratory and the field, with a growing emphasis on integrating agronomic research with fundamental plant science and -omics data in order to understand the molecular mechanisms that underpin key crop traits, variation and performance, as well as to make use of molecular technologies for breeding and other applications (Harfouche et al., 2019; Sperschneider, 2019; Dobrescu et al., 2020; Wang et al., 2020).

Moreover, the last two decades have seen the creation of hybrid research spaces, such as smart glasshouses and digital farm platforms, that utilise new sensing and imaging technologies to capture features of the environment with unprecedented precision and scale (Coppens et al., 2017; Tardieu et al., 2017; Giuffrida et al., 2018). Each of these research spaces hosts different constellations of interdisciplinary work, whose diverse methods and outputs can be challenging to consider as a single body of evidence. Add to these scientific concerns the legal challenges presented by frequent (and frequently unresolved) clashes between different intellectual property regimes. More egregiously, there is a tension between publicly and privately funded research efforts. Much plant research takes place under the auspices of the agrotech industry, whose tendency to keep data in-house, due to its commercial sensitivity, differs substantially from the Open Science ethos characterising much publicly-funded plant science, where large-scale research around model organisms like *Arabidopsis thaliana* resulted in an extensive set of standards, conventions and platforms devoted to effective data sharing and the idea of data as “knowledge commons” (Leonelli, 2016a; Henkhaus et al., 2020). Tensions between competing claims to national sovereignty over biological materials and related data, as well as the jurisdiction of different types of licenses, patenting systems and copyright agreements, further complicate this landscape. Last but not least, at the social end of the spectrum plant-related work involves many contributors beyond professional research circles, including farmers and their communities, breeders, food producers, and policy-makers involved in agricultural policies at the regional and national levels and trade agreements at the international level. These diverse participants tend not to communicate effectively with each other. Differences in skills and goals, social divides, persisting power asymmetries and the sheer quantity of relevant stakeholders make it particularly hard for farmers to provide input and feedback to researchers and policy-makers, and thus to contribute to discussions around what counts as scientific findings, what those may signify for agricultural development within local territories, and what role digital technologies can and should play in land management and food production.

This volume starts from the recognition that *scientific, legal, political and socio-economic challenges such as these are inextricable from each other and have a decisive impact on which plant data get to circulate, to whom and for which purposes*. An immediate implication of this premise is that confronting these challenges requires an awareness of the complex landscape in which they emerge, including some understanding of their historical roots. This volume is intended as a multidisciplinary, transnational entry point to that landscape. It assembles a wide range of practitioners from data science, ethics and the law, history and social studies of science and agronomy, which together represent some of the key initiatives in plant data linkage and curation in the world. The volume thus examines the opportunities and challenges of plant data linkage and re-use as experienced by contributing authors who have spent decades working in this domain. Our goal is to chart and support data exchanges that are not only scientifically and agronomically productive, but also responsive to the social circumstances in which data and plants are collected and used – and in that sense, are both effective and responsible.

In this introduction, we provide essential background to this work. In the next section, we examine the different meanings that the idea of responsible practice can take at the four key sites of plant data governance: the plant environments being documented (the *field*), the infrastructures used to circulate information (the *data*), the entities involved in data governance (the *institutions*) and the variety of expertise and interests involved in plant data work (the *communities*). In Sect. 3, we then outline what we regard as four crucial steps towards achieving responsible data linkage: (1) the building of infrastructures to foster critical data reuse; (2) the development and implementation of transnational legal and institutional frameworks; (3) the formulation of effective ethical guidance and related monitoring systems; and (4) the creation of mechanisms to identify and regularly evaluate assumptions made about agricultural development and the contribution of agricultural science to society, and to consider alternative frameworks. In conclusion, we emphasise the importance of giving equal consideration to these four steps not just in developing but most importantly in maintaining responsible and fruitful practices of plant data linkage in the long term – a crucial factor in making such practices trustworthy and dependable.

2 Dimensions of Responsible Plant Data Governance

There is increasing awareness of the enormous resources and labour required to develop data infrastructures through which data and knowledge about plants can be garnered and harnessed appropriately. These include tools that can foster harmonious data exchange and mining, such as semantic systems, formatting standards, metadata categories and tailored databases. Developing such tools is a technical challenge that has kept thousands of computer, plant and data scientists busy for decades. As many contributions to this volume illustrate (Bertin et al.; Devare et al.; Rawlings and Davey; Pommier et al.; Ostler et al.), such efforts have yielded impressive progress, with substantive innovations emerging to help curate and organise plant data for future re-use. Nevertheless, we remain far from the seamless global systems for data collection and access that were envisaged already at the turn of the last century, when organisations like the League of Nations started to promote systematic efforts to garner and integrate scientifically relevant information from across the world (Hewson, 1999; Edwards, 2010). The vision of all-encompassing automated data analysis linked to the rise of computing in the 1960s and 1970s has not yet materialised, despite the resources devoted to building digitised data infrastructures and the hype surrounding the mining of big data (Williamson et al., 2021).

A key reason for this gap between expectations and reality is that assembling reliable data systems is not only a technical issue, and making plant data amenable to reuse is more than a technical challenge. The creation and curation of interoperable data involves a range of conceptual and social challenges that are inseparable from the technical aspects. For example, in order to make given plant traits amenable to large-scale computational analysis, it is necessary to have suitable labels for the data

clusters relevant to investigating such traits. This requires the development of reliable and standardised trait descriptors, which in turn involves consultations across breeders, farmers, researchers and consumers concerning which traits are most significant for investigation and which labels are most appropriate in defining them – a fraught set of questions to ask within a cross-cultural, multilingual environment plagued by power differentials and inequity between the parties involved (Arnaud et al., 2020; Leonelli, 2022; Curry & Leonelli, 2022). Additionally, analysing data on phenomena ranging from ecological stressors to host-pathogen interactions requires having sufficient metadata about the conditions of origin and the legitimate range of possible uses of such data (Shaw et al., 2020); and linking data from many different sources (whether genomic and experimental data from public or corporate research, knowledge of plant strains and environments held by farmers and breeders, or data related to stored germplasm collections) requires sharing, access and reuse agreements among stakeholders as well as venues in which such agreements can be forged. These are very complex requirements given the diverse regimes of intellectual property, commercial sensitivity, research incentives, cultural ownership and trade to which data are subjected, and the existing tensions around the goals, motivations, and implications of data disclosure and re-use.

All this makes the idea of ‘responsible practice’ in plant data management difficult to understand and operationalise. What does responsibility mean here, given how distributed and diverse plant data stakeholders, contributors, infrastructures and users are?¹ Our starting point in answering this question is to acknowledge that responsibility means different things depending on the setting and goals it needs to serve. Thus, rather than trying to settle on a unique and common definition for this notion, we review four key dimensions of data linkage, and examine what responsible practice may signify within each. These dimensions of data practice also provide the main structure for this volume, which is divided into four parts accordingly.

2.1 The Field: Documenting Variability in Plants and Their Environments

A recurring concern in the management and curation of plant data is the extensive variability encountered both in the plant specimens and in the environments in which they grow, including the intersections of such environments with human

¹In approaching the topic of responsibility in scientific and technical domains, we follow in the path of extensive work undertaken under the aegis of ‘responsible research and innovation’ (e.g. Owen et al., 2012). As presented in this introduction, our framing of issues of responsibility is tailored to the specific issues in the fields of plant and agricultural science. For a related summary of ‘responsible data’ issues in agriculture, see Ferris and Rahman (2016); for considerations of responsible data governance within highly distributed technical systems, see Edwards et al. (2011), Lagoze (2014) and Leonelli (2016b).

communities. It is critical for plant data systems to capture accurate information about which species and varieties are being documented and which seeds are collected, as well as which environmental features are most relevant to plant development and yield. And there is broad agreement on the prominent role that genetic information has come to play in supporting this effort, and therefore on the significance of sharing digital sequence information as a gateway to understanding agrobiodiversity (Morgera et al., 2020). Nevertheless, the variability in the characteristics of plants and their environments (including, crucially, the soil) is extensive and highly dynamic, particularly under conditions of climate change. Moreover, such environmental variability is flanked by variability in the methods and procedures used to generate data and curate relevant materials (such as germplasm), as well as social variability in the preferences, assumptions and conceptual commitments held by data producers and stewards. Settling on data practices and standards to capture such information is a priority and a serious challenge, with important repercussions on the systems used to evaluate performance, productivity and success of agricultural strategies.

When considering the processes involved in extracting data from local fields, crops, seed systems and their environments, the central concern around responsible data practice thus relates to decisions around which kinds of variability need to be reported into data systems. Responsible data linkage involves explicitly asking how different kinds of variability feature in data systems and the ways in which the success of such systems is assessed, and ensuring that the decisions taken in response to this question are regularly scrutinized and reviewed across stakeholders. As the chapters in the first part of this volume make clear, this is hard to implement in practice for a variety of reasons. One is scientific disagreements on how data may be interpreted, which Radick's chapter discusses under the heading of "[Theory-Ladenness as a Problem for Plant Data Linkage](#)" and elegantly exemplifies with reference to the history of Mendelian plant genetics. Another is the cost and technical intricacy of harmonizing various types of environmental data with data about crops, especially considering the evolution of seed trade, intellectual property and public-private relations underpinning modern breeding – as beautifully illustrated, through a narrative spanning the whole of the twentieth century, by Harrison and Caccamo's chapter "[Managing Data in Breeding, Selection and in Practice: A Hundred Year Problem That Requires a Rapid Solution](#)". A third consists of the diverse political conditions under which specific taxonomies of seeds may come to be defined and valued as objects of analysis, which in turn determines the characteristics of related data collections. This is poignantly exemplified by Fullilove's and Alimari's analysis of wheat breeding and preservation projects on the West Bank in their chapter "[Baladi Seeds in the oPt: Populations as Objects of Preservation and Units of Analysis](#)". And last but not least, there are concerns around how the apparatus devised to extract and manage data intersect with breeding practices on the ground, which call for the establishment of effective and mutually respectful dialogue between data linkage experts and those who run field trials and provide key materials and observations. Efforts in this direction are exemplified and discussed by Agbona and colleagues in relation to Root, Tubers and Banana crop breeding programmes in

Africa in the chapter “[Data Management in Multi-Disciplinary African RTB Crop Breeding Programs](#)”.

These challenges are not only about the technical assemblage of data sources, though this is certainly a crucial problem in this domain. Among the broader issues raised by these studies, we find a systematic questioning of the extent to which data management methods focused on digital sequence information can fruitfully serve broader phenotype and environmental datasets; of how breeding strategies are identified and chosen, and with what implications (for example when privileging *ex situ* breeding over *in situ* efforts, as long done by many research institutes around the world; see also Curry, 2017 and Curry’s chapter in this volume); and whether and how data systems can and should pay more attention to marginal environments where uniform crop varieties do not perform consistently, rather than prioritizing data collection on selected high yield varieties.

2.2 The Data: Developing Scalable and Interoperable Infrastructures

These issues become even more pronounced when shifting attention from the field and circumstances of data collection to the nature of the data themselves, and how the characteristics of data affect efforts to develop and link data infrastructures. A key concern in this respect is how to bring data together in the first place. The idea of *integration* is often used to refer to the ability to aggregate and analyze different datasets as if they were a single body of evidence. Yet integration conceptualized in this way is very demanding: it requires making specific choices as to what the best ways to format and visualize data may be, which may be well-suited to the question at hand but not to other forms of data re-purposing; and yet it may be difficult to disaggregate the data once they have been fully integrated. These concerns are the reason why *interoperability* has taken the place of integration as a crucial and potentially more responsible form of data linkage. Interoperable databases are those that enable their users to ask common queries, thereby supporting links between datasets without reducing users’ ability to ask different questions and re-purpose the data accordingly. Interoperability can foster the accountability of data practices, by making it easier to track who has selected and co-analysed which data, from where and how – and thereby being more responsive to the diverse interests and goals of data users. Effective interoperability requires, in turn, at least some level of standardization in both datasets and data infrastructures, which can facilitate common searches and make data comparable in the desired respects.

This is where the desire for interoperable data meets the problem of *scalability*. It is hard enough to set up a data infrastructure able to capture, store and disseminate data obtained from different field trials carried out by a specific institution on a particular crop – such as the UK-based work on wheat documented by Rawlings and Davey’s chapter “[From Farm to FAIR: The Trials of Linking and Sharing Wheat](#)

[Research Data](#)". But as Rawlings and Davey show so effectively in their discussion of the Designing Future Wheat research programme, the thorniest issues emerge when trying to link such data to data obtained on field trials carried out elsewhere or on other crops, or even to other types of data on crops (such as phenotyping or experimental data). The success and scalability of data practices depend on the effectiveness of Field-to-Lab-to-Field cycles, in Rawlings and Davey's words, where those involved in generating, standardizing and interpreting the data have means to interact regularly and give each other feedback.

Of course, data work can be scaled up further by going beyond the field and lab environments to include environmental research and statutory data produced through agronomic governance, as discussed at length in Harrison's and Caccamo's chapter. Yet another way to scale up data practices is to take a longitudinal view of agricultural research, and link data produced in the present with data generated by the many decades of experimentation which preceded the digital era, while also paying attention to how data collected from very long-running experiments should be managed to enhance their usability now and in the future. The chapter "[Linking Legacies: Realising the Potential of the Rothamsted Long-Term Agricultural Experiments](#)" by Ostler and collaborators from Rothamsted Research, one of the longest-running agricultural research stations in the world, closely examines means of facilitating data scalability and interoperability in time as well as in space, and challenges emerging when considering legacy data. These are crucial concerns at a moment where many data infrastructures are set up to serve specific projects through short cycles of funding, leaving the future maintenance of those databases in limbo, and agricultural institutions around the world host precious, non-digital data collections stretching back several decades, whose potential value to plant research is limited by their inaccessibility. Among the solutions developed at Rothamsted to these challenges, including the design of the electronic Rothamsted Archive (e-RA) database, the emphasis placed on *skilled data curation* is particularly notable. As evident in almost all contributions to this volume, data curators play a key role in mediating between the archive and would-be users, bringing expert knowledge of datasets and experimental narratives (i.e. the history and purposes of each experiment) to bear where standardisation alone is insufficient to ensure effective and responsible reuse.

A fruitful way to conceptualize and explore standards for data linkage, and support the development of interoperable systems, is to consider whole *data lifecycles*. This involves rejecting a strict compartmentalization of different types of data practices, such as for instance data production, cleaning, formatting and modelling, and instead understanding such data practices in terms of how they relate to each other within and beyond the world of research (Borgman, 2019; Leonelli, 2019). In their chapter "[Plant Science Data Integration, From Building Community Standards to Defining a Consistent Data Lifecycle](#)", Pommier and colleagues reflect on the ongoing attempt to develop data standards that are meaningful and useful to specific communities of plant researchers, while also taking account of how such standards may support subsequent stages of the data lifecycle in a consistent manner. As they note, data standards are only effective in promoting interoperability as long

as they are successfully implemented, and the conditions of successful implementations prominently include the degree to which standards are tractable, trusted and perceived to be useful among users.

2.3 The Institutions: Overseeing the Dissemination and Use of Plant Data and Materials

Concerns around which kinds of expertise, venues and social arrangements are most appropriate to facilitating data linkage have already repeatedly come to the fore in our discussion, and it is therefore no surprise that the third domain we wish to highlight is that of the institutions responsible for devising and implementing data governance strategies. Responsible practice here includes not only the design of rules and regulations that may support – rather than hinder – data work, but also regular monitoring of the extent to which these systems are being implemented, and most importantly, of their impact on plant research as well as agricultural and food systems. Ultimately, responsibility in this domain means taking ownership of both the positive and negative social consequences of specific data practices, and taking action whenever a given governance method fails to support agricultural and social development. This in turn requires ongoing consideration of what constitutes desirable development, and for whom.

What organizational and governance structures are fit to address such a challenge? Devare and colleagues consider this question in their chapter “[Governing Agricultural Data: Challenges and Recommendations](#)” through a discussion of the forms of leadership, strategy and management required to support data linkage within the CGIAR, a large international organization comprised of 15 agricultural research institutes around the world. The history and current structure of CGIAR effectively exemplifies the opportunities and obstacles created by the requirement to link highly diverse data, coming from culturally, geographically and socio-economically distant communities, in ways that inform agricultural development on a global scale. The central coordination efforts within CGIAR depend on a plethora of other institutions, ranging from the individual CGIAR institutes themselves (each of which has its own governance structure, which is in turn responsive to the specific territory and political situation in their host countries) to the various private and public funders involved in sponsoring projects carried out by CGIAR institutes, the many collaborative networks and consortia set up in relation to specific initiatives and crops, and the international regulations under which this quintessentially transnational work takes place.

Beyond such fragmentation and multiplicity, a central governance challenge for international institutions such as CGIAR is the large *inequity* that characterizes agricultural research across different locations, with many parts of the developing world (and particularly ex-colonies) routinely serving as providers for biological materials and related data and botanical knowledge, and yet not playing an active

role in using the data to produce agricultural innovation (cf. Kloppenburg, 2004; Hayden, 2003; Soto Laveaga, 2009). Unless exploitative practices are appropriately identified and challenged in the course of data work, there is a substantive risk that data linkage strategies may help to further entrench existing systems of unfair data collection and predatory data re-use (Miles, 2019). Fullilove and Alimari's chapter highlights such issues in relation to contemporary seed systems and agricultural development in the occupied Palestinian territories (oPt), thus underscoring how countering in-built inequity and the dominance of the Global North over the agricultural landscape is a priority when seeking to develop responsible systems for data linkage. The chapter "[Digital Sequence Information and Plant Genetic Resources: Global Policy Meets Interoperability](#)" by Manzella and colleagues presents some of the progress made in developing more equitable data systems in tandem with existing policy frameworks for the international governance of plant and agricultural science. These include the systems of access and benefit sharing (ABS) that form a key pillar of the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) and the Nagoya Protocol of the UN Convention on Biological Diversity (CBD). These policy systems and their underpinning legal structure have been challenged in recent years by the high availability of digital sequence information, which has the potential to undermine existing systems of ABS focused on access to germplasm and other biological materials (Morgera et al., 2020). Manzella and colleagues survey the current status of discussions regarding sequence data and ABS policy, focusing on the urgent need to enhance the interoperability of relevant data systems such that the origins and use of sequence data and the status of corresponding biological materials under the ITPGRFA can be easily identified.

Such technical solutions are born of careful consideration of the large political and ethical issues relating to the circulation of plant materials as well as data. The ability to link plant data transnationally is crucial to enhancing biological understanding of crop usage and food systems worldwide, and yet the imaginary of plant genetic resources and related data as 'common goods' remains in tension with national systems of governance for agricultural resources (Bonneuil, 2019). The very idea of (national) *sovereignty* associated to plant materials and data is itself a double-edged sword: it is important to acknowledge and respect, especially given postcolonial legacies of exploitation of specific countries, but it also supports highly restrictive understandings of who may own and use crop data. Responsible data linkage involves tackling these issues through the co-creation of governance and technical systems capable of mediating legal, ethical and social considerations. Kochupillai and Köninger's chapter, "[Creating a Digital Marketplace for Agrobiodiversity and Plant Genetic Sequence Data: Legal and Ethical Considerations of an AI and Blockchain Based Solution](#)", presents an ambitious proposal for how new digital systems could be put in place to overcome some of the current obstacles of supply and demand of agrobiodiversity for research and breeding. Central to their proposals is the need for cutting-edge technical solutions that not only facilitate in situ innovations for farmers as well as researchers, but also respond to current inequities in legal and regulatory regimes. This requires a wholesale

rethinking of key components of contemporary regulation, such as the current dependence of benefit sharing mechanisms on downstream intellectual property rights. The economic implications of such a transformation are vast, both in their consequences for the seed and food markets and in their demands on current investment in data-intensive technologies and related practices.

The relevance of *economic strictures*, and the clash between the need for transformation and the ever more limited resources available for the development and maintenance of reliable data systems, is aptly illustrated by Curry's chapter "[Data, Duplication, and Decentralisation: Gene Bank Management in the 1980s and 1990s](#)". Her analysis of the 'rationalisation' of gene bank collections illustrates a recurring tension between idealised efforts of conservation and reuse of plant-related resources (such as the attempt to assemble comprehensive collections of viable seeds from all over the world) and the lack of the financial and organisational resources required to maintain such plant resources over time. This example shows how policy and organisational solutions implemented on the ground are rarely straightforward responses to data challenges, but are entwined with the need to respond to many competing imperatives, including expectations around what constitutes a profitable investment and the timescale of economic returns. The implications of the political economy of collecting, and the costs involved in long-term maintenance whose economic impact is hard to quantify, must always be borne in mind when evaluating and designing institutional and governance strategies for plant data management and linkage (Strasser, 2019). Whether practitioners explicitly acknowledge it or not, developing long-term data linkage strategies typically involves challenging short-term arguments for predefined and easily quantified sources of economic return, and emphasising instead the diffuse – and even more impactful – ways in which data governance systems may support economic growth and sustainable agricultural development (Leonelli, 2022).

2.4 The Communities: Perspectives from and Accountability to Farmers and Consumers

Perhaps most fundamental and challenging of all is the recognition that plant data – like all other forms of data – have multiple *values* depending on who handles them and for which purposes. Beyond their obvious scientific and commercial value, they may hold affective value, cultural value (if they document knowledge by local communities, for example) and political value (e.g. in disputes over ownership of biological resources). The constellation of relevant values will vary between stakeholders, and there are often tensions between different values held even by single individuals – let alone distributed networks of stakeholders. Recognition and debate around such values is crucial to responsible data practices, which play an essential role in connecting different stakeholders and facilitating communication across communities (Leonelli, 2016b).

When thinking about the governance, circulation and use of plant data, it is crucial to broaden the conversation from the more technical discussion of standards and curation strategies characterising data and plant science circles, and to bring in perspectives from farmers as well as consumers of crops (whether as food, medicine, fuel, fabric or other), and other stakeholders in seed and food systems. It seems trivial to assert that the rights, needs and goals of these communities need to be foregrounded and included in the processes through which infrastructures, governance regimes and policy directions are shaped; and yet, farmers are rarely consulted on and included into the design and governance of data exchange systems. The chapter from Zampati, “[Ethical and Legal Considerations in Smart Farming: A Farmer’s Perspective](#)”, examines the proliferation of ways to extract and monetise data from farmers’ everyday activities, often resulting in exploitative technologies that may benefit the national economy but do not necessarily benefit individual farmers and their communities, and in fact remain unintelligible to farmers and far removed from their sphere of intervention. The chapter presents some models to increase farmer engagement, focusing especially on the adoption of codes of conduct that encourage a dialogue between farmers and the data experts and companies involved in smart farming.

Looking instead at efforts to meaningfully link data from diverse territories and crop varieties with each other, in their chapter “[Communities of Practice in Crop Diversity Management: From Data to Collaborative Governance](#)” Louafi and collaborators provide an example of what they call ‘collaborative governance’, whereby a *community of practice* is constituted to help address both technical and social challenges involved in data linkage. This is a case where the heterogeneity of stakeholders is transformed from a problem into an asset: regular consultation among different experts, including farmers as well as breeders, consumers and data experts from a variety of different territories, becomes a crucial way to understand and manage crop diversity, and thereby build plant data infrastructures that successfully incorporate wide-ranging knowledge sources of relevance to agricultural development. Another great example of a community of practice at work is provided by Rocha Bello Bertin and collaborators, whose chapter discusses the efforts to build such communities by “[The Research Data Alliance Interest Group on Agricultural Data: Supporting a Global Community of Practice](#)”. This volunteer group has spent over a decade on efforts to identify and assemble communities of practice that can support long-term discussions and decisions around the standards and semantics to be used when linking crop data from around the world. Notably, this group has long been open to participation from any relevant stakeholder around the world, and yet – as they observe in the chapter – found several obstacles in integrating wide-ranging expertise and new voices into their work. Being included in data governance efforts often requires some expertise in, and understanding of, existing data systems, as well as the time and resources to find and engage with the right international groups. This is an additional burden on the shoulders of farming communities already under pressure to produce high yield under increasingly competitive and adverse conditions. Engaging in communities of practice can also be slow and sometimes tedious work, replete with discussions over what are

sometimes minute aspects of data curation and standardisation – issues that may matter very little to some of the stakeholders, but crucially affect others.

Questions around the role and incentive structures for communities of data practice parallel long-held debates over the relation between the germplasm acquired from farmers and breeders and the digital data produced by researchers, industry, and governmental institutions. There are sometimes many degrees of separation between the biological materials produced by farmers and the various types of data (molecular as well as administrative and socio-economic) generated by those tasked with analysing and regulating food production and distribution. Given the diverse types of labour and contributions to innovation in such a complex system, it is important to ensure that benefits are equally distributed across the “data chain”, including to farmers and other data providers, rather than being captured by certain end users or those who hold intellectual or other property rights. Equally important is problematising the question of what constitutes a benefit to different stakeholders in the first place, and under which circumstances. The question of adequate and appropriate benefits is one that is hard to address through purely quantitative analysis, and often requires the kind of context-sensitive inquiry that the qualitative social sciences and Science and Technology Studies (STS) are well-placed to carry out. Social scientists are also well-placed to collaborate with both data scientists and farming communities, and thereby help broker conversations and exchanges between different groups.

This is not only an exercise in inclusion for inclusion’s sake. As argued by Radick’s chapter as well as Williamson’s and Leonelli’s, the development of any data system unavoidably involves making strong *conceptual assumptions*, which affect and shape social relations, research goals and even the types of expertise which are regarded as relevant. These assumptions become entrenched into those technical systems and thus increasingly difficult to challenge. At the same time, however, the purpose and reach of those systems continues to change and expand, raising questions as to whether the initial assumptions made when creating those data infrastructures continue to be valid and fruitful. For instance, in their concluding chapter on “[Cultivating Responsible Plant Breeding Strategies: Conceptual and Normative Commitments in Data-Intensive Agriculture](#)”, Williamson and Leonelli discuss how even apparently value-neutral, scientific concepts such as the notion of genetic gain in plant breeding – which is increasingly used as a measure for the productivity of specific crops – can embody a restrictive normative vision for what agricultural development means, how it can be measured and incentivised, and who it is supposed to benefit.

Which criteria are used to single out a desirable plant trait? Are farmers and breeders consulted on which plant trait is most valued by consumers in local markets? Is soil health factored into data systems meant to document field trials, or are the data focusing exclusively on genetic markers for the plant varieties themselves? Asking such questions is a way to critically question received views on the relationship between crops and their biological and social ecosystems, which may be implicitly embedded into data system and linkage tools. Data infrastructures are most often born of the need to compile and circulate an existing dataset, and are thereby often conceptualised as a neutral container – a black box whose only function is to preserve

and spew out data whenever required, and whose functioning should not affect the data and the ways in which they are repurposed. As the chapters in this volume demonstrate, however, there is no such neutrality: rather, data infrastructures are unavoidably value-laden and replete with normative assumptions about what counts as sustainable ways to care for the environment, cultivate crops and produce food. Responsible data practice involves regularly opening and re-ordering the black boxes, checking that their components – including their conceptual apparatus – are fit for their constantly shifting purposes. Ultimately, data linkage systems are systems of relations: taking time to define and regularly re-evaluate what count as relevant *relata*, depending on one's goals, is therefore paramount.

3 Steps Towards Responsible Plant Data Linkage

Who is then responsible and accountable for decisions around data management and the re-use of data, and mistakes or problems associated with such decisions? For example, regarding the allocations of rewards and rights, we might ask who is responsible for “data production” in a given experiment. Is data production the result of growing the plant specimens, selecting strains, designing field trials, adopting novel measurement tools or designing data storage? The answer to this question will determine who is viewed as the legitimate owner of data and who has control over their use. Yet all of these activities have a legitimate claim to being part of data production. Indeed, the chapters of this volume demonstrate the diversity and pervasiveness of responsible practice across the main domains of plant data linkage, which raises urgent questions around the meaning of accountability in such fragmented and distributed systems of knowledge production. All those who participate in plant data analysis – and related benefits and profits – are arguably accountable for their work in some way: their contributions should be evaluated with an eye to their role and consequences within the whole system, and there should be mechanisms to reward good practice and discourage problematic or wrong decisions. However, evaluation of what may constitute responsible practice lags behind. It remains hard to determine what such distributed accountability means in practice; who may be held responsible – and with which implications – when things go wrong; and how to differentiate between human error, system bias and deliberate misuse. In this section, we point to four essential steps towards fostering responsible behaviour while also helping to identify and address problematic data practices.

3.1 Focusing on Critical Data Reuse

A starting point for responsible data linkage is the acknowledgment that the problems of accessing and using data cannot be separated. In other words, Open Data can and should not be a goal in and of itself. Focusing solely or primarily on “putting

data online”, without worrying about who may access such materials, how and for which purposes, is a recipe for disaster. Notably, data linkage makes concerns around data access and re-use inextricable from each other – for while there is no opportunity for linkage without some level of data access, linkage methods unavoidably serve specific expectations of data may be re-purposed. Openness thus needs to be intelligent (Boulton et al., 2012); data infrastructures and tools for data analysis should be developed with at least some awareness of the ways in which data may – or not – be employed in the future, and the types of users who may be involved. Of course, the future of data is never certain or fully predictable, especially in the era of data-driven analysis (Leonelli, 2016a). This does not mean, however, that thoughtful consideration should not be given to the priorities and assumptions built into data linkage systems; in fact, the unpredictability of data use is a key reason to pay close attention to the design, maintenance and broad impact of data linkage systems.

An important step towards refocusing data practices on data reuse is exemplified by the FAIR principles for data management. These principles, now widely recognised worldwide, define effective data sharing as making data Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). This comes with the acknowledgment that Open Data are not always required and never sufficient to guarantee data re-use. Being able to access data that have been badly curated and annotated is often as bad as not having access at all, since data that are badly curated are near-impossible to re-use meaningfully. Furthermore, as is well-acknowledged in the biomedical domain, data can be made available for re-analysis and re-purposing even without direct access: for example, through data mining techniques such as DataShield which facilitate pooled data analysis without sharing individual-level data (Murtagh et al., 2012). The FAIR principles thus took attention away from sheer data access and re-focused instead on *the conditions for* “best data practice”, which in turn involve critically investigating what data exist, whether or not they can or should be accessible, what mechanisms should be used to grant access and how such mechanisms will inform re-use. As part of such efforts, data history (including data provenance as well as the locations, methods and interests of those involved in data processing) is increasingly recognised as essential meta-data that needs to be adequately tracked and documented (Leonelli, 2020). Indeed, within the FAIR framework metadata are arguably more important than data themselves – without appropriate meta-data, data re-use is compromised and the opportunities to re-purpose data are radically restricted, if not altogether eliminated.

As repeatedly noted by our contributors, the FAIR Principles are widely recognised in plant science (Pommier et al., 2019; Reiser et al., 2018) and increasingly built into data collection at source, for example through the creation and use of digital fieldbooks that facilitate the standardisation and semantic interoperability of field data collection (e.g. Rife & Poland, 2014). They are also recognised at the infrastructural level, exemplified by the incorporation of FAIR data metrics into the CGIAR Big Data Platform’s GARDIAN search tool, and there are now dedicated tools to assist in the management and deposition of FAIR data and metadata, notably the Collaborative Open Plant Omics (COPO) platform (Shaw et al., 2020). The extensive implementation of FAIR is a big step forward in the development of plant

data infrastructures that facilitate extensive and responsible linkage, not least for recognising that data access should be carefully monitored and regulated (as often stressed by FAIR data proponents: “data should be as open as possible, as closed as necessary”). This is crucial to enable a more critical and nuanced understanding of the multiple social contexts of data sharing processes, and the potential implications of granting data access in the case of sensitive data. However, as we will see in the next section, this framework does not go far enough, and indeed does not directly include attention to ethical aspects such as equity and fairness in the provenance, ownership and distribution of data resources.

3.2 Encouraging Multiple Forms of Transnational Data Governance

The regulatory framework for plant data work is as yet vague and unclear, with few (if any) existing international agreements concerning the goals, rewards, responsibilities and rights pertaining to the generation, circulation and use of digital plant data. This situation contrasts stridently with the biomedical field, where such agreements have been at the centre of developments in genomics (Maxson-Jones et al., 2018) and the set-up of structures and regimes of data governance for health-related data (Hilgartner, 2017). This arguably owes much to the distinctive risks associated with the category of “personal data” about individual patients, a category which, with the exception of data documenting the socio-economic status of individual farmers, is of less relevance to the plant sciences. And yet, the dissemination and linkage of plant data bears its own social and ethical risks. First, as we discussed above, data sharing across countries remains an underregulated and yet sensitive matter, where data produced in the Global South is systematically harnessed and profitably re-used in the Global North and yet such appropriation often happens without proper attribution and compensation.² Second, large agrotech corporations dominate plant data production and re-use (including through remote sensing technologies incorporated within agricultural machinery) in ways that are rarely transparent and well-aligned with equivalent efforts in the public domain (Shiva, 2016; Fullilove, 2017; Miles, 2019). This makes dialogue around regulation, technical standards and socio-economic implications of data re-use even harder, as there is no overarching sense of the amount, variety and nature of existing data of relevance. Third, the commercial value and cultural capital associated with plant data – and particularly data about indigenous crops – is well-recognised by most countries/governments as a national resource, and yet there is little clarity around whether the deployment of such resource does (or should) reflect national interests, and how this

²This phenomenon can be construed as the digital equivalent to bioprospecting: a “digital feudalism” building on centuries-long exploitation and discrimination built into the food production system (Scott, 1998; Hayden, 2003; Mazzucato, 2019).

sits vis-à-vis the conception of plant knowledge as a global common good (Kloppenborg, 2004; Krige, 2022).

All this indicates that technical means to enable plant data linkage need to be accompanied by an effective system of *transnational data governance*, comprising both the norms and the infrastructure needed to share and re-use data adequately and responsibly. Most contributions to this volume can be read as working towards this goal, whether by developing sharing standards, legal frameworks, governance venues, ethical norms or physical tools. The diversity of such work shows how benefits to be distributed across the data chain include economic gain (as in the proposal of a blockchain solution in Kochupillai and Köninger's chapter) as well as opportunities for shaping data work (through the communities of practices discussed by Rocha Bello Bertin, Louafi and their colleagues in their respective chapters) and be appropriately rewarded for that effort through proper acknowledgment, as fostered by current FAO efforts discussed by Manzella's chapter, and the tracking of data provenance promoted by GODAN, as exemplified in Zampati's chapter. What such suggestions will involve in terms of legal frameworks both nationally and internationally is a crucial problem whose resolution goes well beyond the scope of this volume, but which we hope these contributions may help to inform – particularly by highlighting the diverse levels of governance involved in making data linkage work for users (see also Welch et al., 2021), and fostering a better integration of socio-political concerns into technical efforts to develop plant data infrastructures.

3.3 Developing Guidance in Tandem with Incentives and Monitoring Systems

How to achieve such integration? Alongside the infrastructural work to facilitate critical data reuse and the regulatory work to ensure the legality of data exchanges especially at the international level, there have been increasing efforts to ensure that ethical considerations are built into the design and use of data infrastructures. One mechanism for this has been the creation of additional guiding principles, complementary to FAIR, that are focused on ethical issues.

One such set of principles are the TRUST principles proposed by the Research Data Alliance, which stand for: Transparency, that is the need to make data operations as easy as possible to understand and scrutinize; Responsibility; User Focus, which involves prioritising the needs, skills and concerns of users over the wishes of infrastructure developers; Sustainability, which implies attention to the long-term prospects and environmental impact of the infrastructure; and Technology, that is the importance of keeping an infrastructure up-to-date with evolving software and hardware requirements (Lin et al., 2020). Informed by such principles, there is an emerging trend in broader data science towards public and collective use of knowledge and infrastructures, with a number of data initiatives built with these values at their core (including for instance the Ada Lovelace Institute in London, the Centre

for Technomoral Futures at the University of Edinburgh, the research line on Digital Infrastructures for the Public Interest at Stanford PACS, the PublicSpace coalition in Amsterdam and the Institute for Digital Public Infrastructure at UMass Amherst, to mention just a few). This trend is also visible in programmes that prioritise a responsible approach to research and innovation or to human centric and trustworthy data technology, prominently fostered by the European Commission. It is high time that such approaches are explicitly extended to the plant and agricultural domain, as our Exeter Centre for the Study of Life Sciences at the University of Exeter is attempting to do.

Another important development are the CARE Principles for Indigenous Data Governance, which were produced by the Global Indigenous Data Alliance in consultation with a very wide range of data subjects, producers and users. The CARE Principles draw attention to the implications of open data sharing for indigenous and other communities from whom data may be extracted,³ by focusing on four key issues: (1) equitable distribution of Collective Benefits, that is of evaluating the impact of a given data intervention on groups and communities and ensuring that this impact is positive; (2) the recognition of communities' own Authority to Control, which points to the necessity to distribute power and control over the data across the stakeholders involved, rather than placing all control in the hands of one party (especially if this party consists of digital platforms or specific data users); (3) the Responsibility of researchers to communities, which involves the need to clearly acknowledge who is being held responsible when data work goes wrong; and (4) the foregrounding of Ethics at all stages of the data life cycle, which is a broad invitation to monitor the social and moral implication of any kind of data work.

Both TRUST and CARE principles are part of multiple efforts to introduce reflection on wider obligations and responsibilities into the workings of a given data infrastructure. All too often however the nature of such reflection and any changes resulting from it are left open to actors' own judgement and rely on voluntary adherence. This reflects concerns in the field of AI about "ethics washing" through the creation of sets of principles or guidelines that co-opt the flag of ethics but potentially do little to actually change how tech companies use data, as well as subsequent counter-critiques of "ethics bashing" (Bietti, 2020). As Kind (2020) has noted, moving beyond ethics washing and bashing requires treating the implementation of ethical principles not just as a narrow technical matter but as a socio-technical one that involves addressing local practice and organisation. Hence what we wish to highlight here is not only the significance of such ethical frameworks for future data work, but also the critical role of systems of incentives and monitoring in making it possible to concretely implement these frameworks. For instance, CARE principles need to be complemented by data labels and validation systems that help certify and monitor adherence to such principles.

A great example is provided by the Traditional Knowledge and Biocultural Labels Initiatives, which "allow communities to express local and specific conditions for

³<https://www.gida-global.org/care>

sharing and engaging in future research and relationships in ways that are consistent with already existing community rules, governance and protocols for using, sharing and circulating knowledge and data” (Liggins et al., 2021). The Biocultural Labels focus specifically on the handling of plant genetic resources derived from crops samples associated to traditional knowledge. These labelling systems have been devised by a consortium of researchers working closely with traditional communities in New Zealand as well as representative bodies for Indigenous Communities around the world, such as the Indigenous Data Sovereignty movement (Hudson et al., 2020). Having been successfully trialled within individual projects and specific collections, they are now being considered for adoption by several large data infrastructures around the world. Such an initiative is very important to data linkage initiatives relating to food and agricultural research, especially given the lack of international agreement on whether and how to govern data sharing through the framework of the International Treaty on Plant Genetic Resources for Food and Agriculture – whose Access and Benefit Sharing mechanisms do not include clear instructions on the status of digital sequence data (Aubry, 2019).

The case of TK and Biocultural Labels shows not only the significance of governance mechanisms to concretely implement principles such as CARE and TRUST, but also – going back to the previous section – the multiplicity of forms of governance required for such implementation. These include large-scale efforts from national governments, prominent research funders, corporations and international organisations such as the Food and Agricultural Organisation and the Convention for Biological Diversity, all of whom can consider mandating the use of these kinds of labels within the data infrastructures and policies that they support; as well as small-scale efforts such as individual projects, research centres and universities, whose reach may be limited but which are much closer to the data practices of interest on the ground.

3.4 Considering Alternatives

The final point we want to highlight is that developing responsible and effective data linkage systems requires bringing infrastructural and ethical strategies in line with the conceptual and normative dimensions of scientific and agricultural practice, including the ways in which both agricultural development and data-intensive research are framed. Major global challenges such as climate change, which require significant rethinking of large-scale systems, cannot be tackled without addressing the conceptual underpinnings of those systems and their implications. This in turn involves identifying the imaginaries of agricultural development and related data usage that are instantiated within existing systems, and asking what alternative ways of constructing and understanding the world could look like, what difference they would make to the principles and values supported by contemporary data linkage infrastructure, and what would be the technical implications and strategies involved in implementing such alternative frameworks.

The example of accelerating genetic gain in plant breeding, discussed at length in our own chapter at the end of the volume, is a case in point. The use of genetic gain as a key indicator for agricultural development needs to be situated in relation to the legacy of the Green Revolution, including the tendency to prioritize increased selection efficiency and breeding outputs over the extent to which diverse preferences, practices and contributions can be built into data practices that inform plant breeding. This generated a trend in twenty-first century science-led agriculture towards conceptualising the growth of molecular breeding and climate-smart agriculture as unrelated – or even antithetical – to farmer engagement and participatory methodologies. As we already discussed, this need not be the case; and yet such a conceptual commitment has severe implications for what responsible data practice is taken to be, and by whom. Responding to a wider set of gendered and other agricultural needs in diverse environments, for instance, will require data mining germplasm collections to spotlight non-elite materials that contain traits of potentially greater relevance to these needs, and then dedicating significant pre-breeding efforts to adapt this material such that they too can benefit from more intensive population improvement (cf. Fadda et al., 2020). When taking such an approach, concepts such as genetic gain may well retain an important role, but may not necessarily feature as a central priority around which all other activity is organised. There remains substantial scope for data-intensive breeding in the service of agricultural development and gender equality, without necessarily structuring major breeding decisions around an algorithmic rationality that conceptualises decision-making as a comparison of metric values.

Similarly, efforts such as the TK and Biocultural Labels are associated with a reconceptualization of the very workflow underpinning data-intensive methods, which challenge the idea of data as raw materials from which knowledge can be extracted through a linear process of analysis and interpretation, and instead support a cyclical understanding of how data are generated and used, with multiple feedback loops between data subjects, data collectors, data stewards and data users. As Devare and colleagues also point out in their chapter, considering a variety of perspectives on how data are used in research, and which workflows can best support the production of reliable knowledge, is a fundamental part of data linkage efforts. Conceptual commitments made in data science, plant breeding and agriculture typically structure – and constrain – the uses of plant data, their paths of travel, and the choice of participants (and related types of expertise) in data collection, circulation and use. When addressing responsible data linkage in the plant and agricultural sciences, it is therefore necessary to consider how different visions of data use may be amenable to achieving different goals, be they economic development, equality of participation or justice in food production systems.

4 Conclusion: Training for the Future

We have reviewed how transnational plant data circulation and re-use is subject to countless constraints and strictures from a variety of perspectives and levels of governance and monitoring. Far from being discouraging, acknowledging these

constraints should foster an imagination of what may constitute socially responsive, sustainable data linkage systems, and an alertness to the variety of conceptual underpinnings such systems could have (think about the visionary quality of the European Open Science Cloud, whose attempt to federate existing research data infrastructures across Europe constitutes an unprecedented feat of data linkage within a highly disruptive and at times openly hostile political and economic environment). At a moment of enormous technological, social and geo-political transformation, it is particularly important to challenge long-held assessments of the impact of structural constraints on available courses of action. This is especially important since, at a practical level, the space to consider such alternative conceptualisations has radically shrunk in agricultural research within the last few decades, as seen in the relative decline of participatory methodologies.

A key tool to push this forward is education. We argued that the current data-intensive model of agricultural research and development is predicated on a distinctive set of conceptual and normative visions for agriculture, and that multiple forms of governance need to be implemented in order to enact responsible data linkage practices. By and large, however, neither scientists nor other stakeholders are trained to identify and evaluate such assumptions or to consider their implementation across different technical, social and political contexts. And yet the importance of training tools and programmes for data scientists, farmers, breeders, researchers working in this space – as well as policy-makers and businesses – was already evident during the Green Revolution, where training programmes such as those devised by the CGIAR centres were very effective in furthering a specific understanding of agricultural development and its implementation on the ground. What would it take to operate at the same scale in the realm of data? Who would be responsible for such training, to guarantee that responsible data practice sits at its core? Should industrial and corporate efforts incorporate these forms of education, and how? And can this be achieved without an acritical commitment to exclusionary approaches to genetic conservation and agricultural development? This volume does not provide exhaustive answers to these questions, but it is our hope that readers will be convinced of the significance of querying what constitutes responsible data linkage in the first place, and take inspiration from the multiple efforts described by our contributors in devising ever more data infrastructures and data sharing solutions to foster sustainable agriculture and a healthier planet.

References

- Arnaud, E., Laporte, M. A., Kim, S., et al. (2020). The ontologies community of practice: An initiative by the CGIAR platform for big data in agriculture. *Patterns*, 1, 100105. [https://www.cell.com/patterns/pdf/S2666-3899\(20\)30139-2.pdf](https://www.cell.com/patterns/pdf/S2666-3899(20)30139-2.pdf)
- Aubry, S. (2019). The future of digital sequence information for plant genetic resources for food and agriculture. *Frontiers in Plant Science*, 10, 1046. <https://doi.org/10.3389/fpls.2019.01046>
- Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. *Proceedings to ACM FAT* Conference (FAT* 2020)*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3513182

- Bonneuil, C. (2019). Seeing nature as a ‘universal store of genes’: How biological diversity became ‘genetic resources’, 1890-1940. *Studies in History and Philosophy of Biology and Biomedical Science*, 75, 1–14. <https://doi.org/10.1016/j.shpsc.2018.12.002>
- Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.9a36bdb6>
- Boulton, G., Campbell, P., Collins, N., Hall, W., Elias, P., Laurie, G., O’Neill, O., Rawlins, M., Thornton, J., Vallance, P., & Walport, M. (2012). *Science as an open enterprise* Royal Society Science Policy Centre report 02/12. The Royal Society.
- Coppens, F., Wuyts, N., Inzé, D., & Dhondt, S. (2017). Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Current Opinion in Systems Biology*, 4, 58–63. <https://doi.org/10.1016/j.coisb.2017.07.002>
- Curry, H. (2017). From working collections to the World Germplasm Project: Agricultural modernization and genetic conservation at the Rockefeller Foundation. *History and Philosophy of the Life Sciences*, 39, 5. <https://doi.org/10.1007/s40656-017-0131-8>
- Curry, H., & Leonelli, S. (2022). Describing crops in the CGIAR era. In H. Curry & T. Lorek (Eds.), *Research as Development: Historical Perspectives on Agricultural Science and International Aid in the CGIAR Era*. Cambridge University Press.
- Dobrescu, A., Giuffrida, M. V., & Tsaftaris, S. A. (2020). Doing more with less: A multitask deep learning approach in plant phenotyping. *Frontiers in Plant Science*. <https://doi.org/10.3389/fpls.2020.00141>
- Edwards, P. N. (2010). *A Vast Machine: Computer Models, Climate Data, and The Politics of Global Warming*. MIT Press.
- 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.
- Fadda, C., Mengistu, D. K., Kidane, Y. G., Dell’Acqua, M., Pe, M. E., & van Etten, J. (2020). Integrating conventional and participatory crop improvement for smallholder agriculture using the seeds for needs approach: A review. *Frontiers in Plant Science*, 11, 559515. <https://doi.org/10.3389/fpls.2020.559515>
- Ferris, L., & Rahman, Z. (2016). *Responsible data in agriculture*. Global Open Data in Agriculture & CABI.
- Fullilove, C. (2017). *The profit of the earth: The global seeds of American agriculture*. University of Chicago Press.
- Giuffrida, M. V., Chen, F., Scharr, H., & Tsaftaris, S. A. (2018). Citizen crowds and experts: Observer variability in image-based plant phenotyping. *Plant Methods*, 14, 12. <https://doi.org/10.1186/s13007-018-0278-7>
- Harfouche, A. L., Jacobson, D. A., Kainer, D., et al. (2019). Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends in Biotechnology*, 37(11), 1217–1235. <https://doi.org/10.1016/j.tibtech.2019.05.007>
- Hayden, C. (2003). *When Nature Goes Public: The Making and Unmaking of Bioprospecting in Mexico*. Princeton University Press.
- Henkhaus, N., Bartlett, M., Gang, D., et al. (2020). Plant science decadal vision 2020-2030: Reimagining the potential of plants for a healthy and sustainable future. *Plant Direct*, 4(8), e00252. <https://doi.org/10.1002/pld3.252>
- Hewson, M. (1999). Did global governance create informational globalism? In M. Hewson & T. J. Sinclair (Eds.), *Approaches to Global Governance Theory*. State University of New York Press.
- Hilgartner, S. (2017). *Reordering life: Knowledge and control in the genomics revolution*. MIT Press.
- Hudson, M. N. A., Garrison, R. S., et al. (2020). Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nature Reviews. Genetics*, 21, 377–384. <https://doi.org/10.1038/s41576-020-0228-x>
- Kind, C. (2020, August 23). *The term ‘ethical AI’ is finally starting to mean something*. Venture Beat. <https://venturebeat.com/2020/08/23/the-term-ethical-ai-is-finally-starting-to-mean-something/>

- Kloppenborg, J. R. (2004). *First the seed: The political economy of plant biotechnology, 1492–2000* (2nd ed.). University of Wisconsin Press.
- Krige, J. (Ed.). (2022). *Knowledge flows in a global age: A transnational approach*. University of Chicago Press.
- Lagoze, C. (2014). Big Data, data integrity, and the fracturing of the control zone. *Big Data & Society*, 1(2), 2053951714558281.
- Leonelli, S. (2016a). *Data-centric biology: A philosophical study*. Chicago University Press.
- Leonelli, S. (2016b). Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production. *Philosophical Transactions of the Royal Society: Part A*, 374, 20160122. <https://doi.org/10.1098/rsta.2016.0122>
- Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualising data in data science. *Harvard Data Science Review*, 1.1. <https://doi.org/10.1162/99608f92.17405bb6>
- Leonelli, S. (2020). Learning from data journeys. In S. Leonelli & Tempini (Eds.), *Data journeys in the sciences*. Springer. Open Access. <https://www.springer.com/gp/book/9783030371760>
- Leonelli, S. (2022). How data cross borders: Globalising plant knowledge through transnational data management and its epistemic economy. In J. Krige (Ed.), *Knowledge flows in a global age: A transnational approach*. University of Chicago Press.
- Liggins, L., Hudson, M., & Anderson, J. (2021). Creating space for Indigenous perspectives on access and benefit-sharing: Encouraging researcher use of the local contexts notices. *Molecular Ecology*, 30(11), 2477–2482. <https://doi.org/10.1111/mec.15918>
- Lin, D., Crabtree, J., Dillo, I., et al. (2020). The TRUST principles for digital repositories. *Nature Scientific Data*, 7, 144. <https://doi.org/10.1038/s41597-020-0486-7>
- Maxson Jones, K., Ankeny, R. A., & Cook-Deegan, R. (2018). The Bermuda Triangle: The pragmatics, policies, and principles for data sharing in the history of the human genome project. *Journal of the History of Biology*, 51, 693–805. <https://doi.org/10.1007/s10739-018-9538-7>
- Mazzucato, M. (2019, October 2). *Preventing digital feudalism*. Project Syndicate. Available at <https://www.project-syndicate.org/commentary/platform-economy-digital-feudalism-by-mariana-mazzucato-2019-10>
- Miles, C. (2019). The combine will tell the truth: On precision agriculture and algorithmic rationality. *Big Data & Society*, 6(1). <https://doi.org/10.1177/2053951719849444>
- Morgera, E., Switzer, S., & Geelhoed, M. (2020). *Possible ways to address digital sequence information – Legal and policy aspects*. Consultancy study for the European Commission. Strathclyde Centre for Environmental Law and Governance.
- Murtagh, M. J., Demir, I., Jenkins, K. N., et al. (2012). Securing the data economy: Translating privacy and enacting security in the development of DataSHIELD. *Public Health Genomics*, 15, 243–253. <https://doi.org/10.1159/000336673>
- Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy*, 39, 751–760.
- Pommier, C., Michotey, C., Cornut, G., et al. (2019). Applying FAIR principles to plant phenotypic data management in GnpIS. *Plant Phenomics*, 2019, 1671403. <https://doi.org/10.34133/2019/1671403>
- Reiser, L., Harper, L., Freeling, M., Han, B., & Luan, S. (2018). FAIR: A call to make published data more findable, accessible, interoperable and reusable. *Molecular Plant*, 11, 1105–1108. <https://doi.org/10.1016/j.molp.2018.07.005>
- Ribaut, J.-M., & Ragot, M. (2019). Modernising breeding for orphan crops: tools, methodologies, and beyond. *Planta*, 250, 971–977. <https://doi.org/10.1007/s00425-019-03200-8>
- Rife, T. W., & Poland, J. A. (2014). Field book: An open-source application for field data collection on android. *Crop Science*, 54, 1624–1627. <https://doi.org/10.2135/cropsci2013.08.0579>
- Scott, J. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.
- Shaw, F., Etuk, A., Minotto, A., et al. (2020). COPO: A metadata platform for brokering FAIR data in the life sciences. *F1000 Research*. [Version 1; peer review: 1 approved, 1 approved with reservations], 9, 495. <https://doi.org/10.12688/f1000research.23889.1>

- Shiva, V. (2016). *Seed Sovereignty, Food Security*. North Atlantic Books.
- Soto Laveaga, G. (2009). *Jungle Laboratories: Mexican Peasants, National Projects, and The Making of The Pill*. Duke University Press.
- Sperschneider, J. (2019). Machine learning in plant-pathogen interactions: Empowering biological predictions from field scale to genome scale. *The New Phytologist*, 228(1), 35–41. <https://doi.org/10.1111/nph.15771>
- Strasser, B. (2019). *Collecting experiments: Making big data biology*. University of Chicago Press.
- Tardieu, F., Cabrera-Bosquet, L., Pridmore, T., & Bennett, M. (2017). Plant phenomics, from sensors to knowledge. *Current Biology*, 27, R770–R783. <https://doi.org/10.1016/j.cub.2017.05.055>
- Wang, H., Cimen, E., Singh, N., & Buckler, E. (2020). Deep learning for plant genomics and crop improvements. *Current Opinion in Plant Biology*, 54, 34–41. <https://doi.org/10.1016/j.pbi.2019.12.010>
- Welch, E., Louafi, S., Carroll, S.R. et al. (2021). *Post COVID-19 implications on genetic diversity and genomics research & innovation: A call for governance and research capacity*. White Paper (released 6 July 2021). <http://www.fao.org/3/cb5573en/cb5573en.pdf>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Nature Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Williamson, H. . F., Brettschneider, J., Caccamo, M., et al. (2021). Data management challenges for artificial intelligence in plant and agricultural research. *F1000*. [Version 1; peer review: 1 approved with reservations]. <https://doi.org/10.12688/f1000research.52204.1>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

